An LPC Analysis and Synthesis System for Arabic Speech

by

Gaswarah Muhammad Dhiyab Abu-Askar

A Thesis Presented to the

FACULTY OF THE COLLEGE OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

ELECTRICAL ENGINEERING

January, 1989

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

AN LPC ANALYSIS AND SYNTHESIS SYSTEM FOR ARABIC SPEECH

BY

Gaswarah Muhammad Dhiyab Abu-Askar

A Thesis Presented to the

FACULTY OF THE COLLEGE OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

ELECTRICAL ENGINEERING

JANUARY 1989

KING FAHD UNIVERSITY OF PETROLEUM AND MINERALS
DHAHRAN 31261, SAUDI ARABIA

COLLEGE OF GRADUATE STUDIES

This thesis, written by GASWARAH MUHAMMAD ABU-ASKAR under the direction of his Thesis Advisor and approved by his Thesis Committee, has been presented to and accepted by the Dean of the College of Graduate studies, in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE in ELECTRICAL ENGINEERING.

THESIS COMMITTEE

Thesis Advisor

Md. Shafiq Ahmad.
Member

Member

Department Chairman

Dean, College of Graduate Studies

Jan 10, 1989
Date

to my mother, father, sisters, brothers and wife....

ACKNOWLEDGEMENT

Acknowledgement is due to King Fahd University of Petroleum and Minerals for support of this research.

I would like to express my appreciation to Dr. Suleymen Penbeci, my principal advisor, for his careful guidance and encouragement through this research.

I am also grateful to other committee members Dr. Muhammad S. Ahmed and Dr. Ubaid M. Al-Saggaf for their useful and valuable suggestions and discussions.

خلاصة الرسالة

اسم الطالب : قسورة محمد أبو عسكر

عنوان الدراسة : بناء نظام لتحليل وتركيب الكلام العربي باستخدام معاملات التنبؤ الخطي

التخصـــــص : هندسة كهربائية

تاريخ الدرجـة : يناير ١٩٨٩م

يعتمد تطوير أنظمة معالجة الكلام بواسطة الحاسب الالي في كثير من الأحيان على طبيعة اللغة المستخدمة.

ولا يشترط للنظام المطور للغة معينة ان يكون مناسبا للغة أخرى. فمثلا: النظام المطور لتحليل وتركيب الكلام للغة الانجليزية لا يناسب لتحليل وتركيب الكلام للغة العربية وذلك للفروقات الواضحة بين اللغتين.

يقدم هذا البحث بناء نظام لتحليل وتركيب الكلام للغة العربية باستخدام معاملات التنبؤ الخطي. وقد تم اختبار نظام مصمم لتحليل وتركيب اللغة الانجليزية لتحليل وتركيب الكلام باللغة العربية فتبين انه غير مناسب. وبالتالي تم تعديل هذا النظام كي يوافق الكلام باللغة العربية بإيجاد القيم المناسبة لطول إطار التحليل ، و عدد أقطاب مصفّيات التحليل و التركيب ودالة إثارة مصفّى التركيب.

درجة الماجستير في العلوم
جامعة الملك فهد للبترول والمعادن
الظهران - المملكة العربية السعودية
يناير ١٩٨٩م

ABSTRACT

Gaswarah Muhammad Abu-Askar

AN LPC ANALYSIS AND SYNTHESIS SYSTEM FOR ARABIC SPEECH

Major Field : Electrical Engineering

January 1989

In many cases the development of a computer speech processing system depends on the natural language used. Hence, systems developed for native speakers of one language may not be adaptable to the native speakers of another language. Most of the research work in speech processing has been mainly for English language. However, few research work was done for Arabic language. The necessity of extensive research work for Arabic was expressed in literature since the properties of Arabic language differ from that of English language.

In this work, a linear predictive coding (LPC) analysis and synthesis system is implemented and tested for the suitable conditions of analysis and synthesis processes to accommodate Arabic language specifications. An analysis and synthesis system developed which was designed with certain conditions to produce good quality synthesized English speech was considered and tested to synthesize Arabic speech. The results were not satisfactory. Accordingly the system was remodeled to produce good quality Arabic speech. The tested conditions were the analysis frame length, number of poles of the analysis and synthesis filters and the driving function for the synthesis filter.

MASTER OF SCIENCE

KING FAHD UNIVERSITY OF PETROLEUM AND MINERALS
DHAHRAN, SAUDI ARABIA

January 1989

CONTENTS

Abstract vi

Chapter I: Background and Literature Review 1

Introduction 1
Speech Synthesis 2
Speech Synthesis Techniques 2
The Channel Vocoder 3
Formant Synthesis Technique 3
Articulatory Synthesis Technique 4
LPC Synthesis Technique 5
Types of Speech Synthesis Methods 6
Applications of Speech Synthesis 6
Properties of Arabic Speech 7
Thesis Motivation 11

Chapter II: Speech Generation 12

Introduction 12
Human Vocal System 12
Speech Production Mechanism 15
Voiced sounds 15
Unvoiced Sounds 16
Speech Production Model 17
Mathematical Model of The Vocal Tract System 17
All-Pole Model of Vocal Tract 20
Linear Prediction of Speech 22
Linear Prediction Coefficient Evaluations 28

Chapter III: LPC Speech Analysis and Synthesis System 31

Introduction 31
Speech Analysis 31
LPC Coefficients Evaluation 32
Sampling Frequency 32
Order of the Filter 33
Frame Length 33
Windowing 34
Preemphasis 36
Pitch Detection and Voiced/Unvoiced Decision 36
ML Algorithm for Pitch Period Detection 38
AMDF Algorithm for Pitch Detection 39

Speech Synthesis 39
Synthesis Filter Structure 39
Excitation and Synthesis Matching 41
Post-emphasis 44

Chapter IV: Implementation and Testing of The System 45

Introduction 45
Data Acquisition 45
Analog Data Collection 45
Digitization and Data Transfer 46
Software Implementation of The System 46
Testing The System Parameters 50
Analysis Parameters 53
Frame Length 53
Number of Poles 54
Synthesis Parameters 61
Driving Function 61
The Selected System Parameters 66
Discussion 74

Chapter V: Conclusion 76

Summary 76
Recommendation for Further Work 77

Appendix A: programs 79

Main program 80
Autocorrelation Subroutine for Calculating The LPC's Coefficients . . 86
AMDF Pitch Detection Program 88
Synthesis Subroutine 93

Appendix B: specification of the equipments 101

References 102

FIGURES

1.1 Arabic vowel triangle. 9

1.2 English vowel triangle. 10

2.1 Human vocal system. 13

2.2 Nasal coupling to the vocal tract. 14

2.3 Speech production model. 18

2.4 Concatenation of (N = 7) lossless tubes of equal length. 19

2.5 Block diagram for speech production mechanism 21

2.6 Linear prediction model in time domain. 24

2.7 Linear prediction model in frequency domain. 25

3.1 Hamming window in time domain. 35

3.2 Speech synthesis scheme. 40

3.3 A two-multiplier lattice filter. 42

4.1 Data acquisition setup 47

4.2 Block diagram of the analysis mode 48

4.3 Block diagram of the synthesis mode 49

4.4 Questionnaire 52

4.5 Spectrographs of one sample set of the first test. 57

4.6 Spectrographs of one sample set of the second test. 60

4.7 Train of impulses driving function 62

4.8 First suggested driving function 63

4.9 Second suggested driving function 64

4.10 Third suggested driving function 65

4.11 Spectrographs of one sample set of the third test. 69

4.12 Original speech and synthesized speech using the selected 71

4.13 Original speech and synthesized speech using the selected 72

4.14 Original speech and synthesized speech using the selected 73

TABLES

4.1 Frame length testing results 55

4.2 Number of poles testing results 58

4.3 Driving function testing results 67

4.4 The average results of the tested parameters. 70

Chapter I
BACKGROUND AND LITERATURE REVIEW

1.1 Introduction

The interest in man-machine communication by voice goes back to the eighteenth century when men first started trying to produce some kind of sounds from primary mechanical devices. However, recent advances in computer technology, and the simultaneous advances in the digital signal processing theory, phonetics, acoustics and artificial intelligence had a very profound effect on speech research. Consequently, an impressive progress has been made in man-machine voice communication [1].

The research on man-machine voice communication can be divided into three categories:

1. Machine speech recognition.
2. Speaker identification.
3. Machine speech response.

The first category, machine speech recognition, deals with techniques that make the machine able to "understand" the input speech. The second category, speaker identification, deals with the ability of the machine to identify a specific speaker who gives a command to it. The last category, machine speech response, deals with the situation in which the machine "replies" in a speech form response to or according to a specific set of conditions in certain applications. Speech synthesis techniques are used for this machine response.

1.2 Speech Synthesis

Two different approaches to the design of speech synthesizers had become manifest by the year 1960 [2]. In one approach, the speech production mechanism itself is modeled by considering physiological details. This approach is called system model. In the other, the speech signal is modeled using whatever techniques convenient to be adopted. This method is called signal model. Signal models are easier to design and construct. This is because it is only necessary to model the effects of the articulation, rather than the mechanism as in system model. On the other hand, system modeling or articulatory synthesis is very difficult to achieve and the most complete models are computational analogues which run as a program on digital computers [2]. The following is a brief review of speech synthesis techniques and types.

1.2.1 Speech Synthesis Techniques

From the literature two major types of speech synthesis techniques can be distinguished

1. Synthesis using pre-recorded speech elements (words, diphones, allophones). These techniques use digitized speech which range from the Hi-Fi 128 kbits/s to the PCM (pulse code modulation) 64 kbit/s down to delta modulation speech at 16 kbits/s.
2. Synthesis by re-constructing speech using parameters of source/filter model which can be realized in the frequency or time domain. The bit rate involved in these techniques ranges from 50 bit/s for articulatory model to 2400 bit/s for linear predictive coding (LPC), up to channel vocoder at 9600 bit/s [3].

The source/filter speech synthesis techniques are used in systems in which low bit rate is desired for transmission and storage. In the following a brief review of source/filter speech synthesis is given.

1.2.1.1 The Channel Vocoder

The principle behind this technique is, on one hand, to find the power spectrum $|S(f)|^2$ of the speech signal as seen by an appropriate temporal window. This spectrum is divided into 12 to 24 bands. The evolution, with time, of the power in each band is considered as one parameter, which means that we can have from 12 to 24 such parameters depending on the quality desired. On the other hand, the fundamental frequency is detected when the type of the source excitation is voiced. We also have to decide whether the sound is voiced or unvoiced. The typical bit rate needed to control a channel vocoder synthesizer ranges from 2400 to 9600 bits/s [3].

The channel vocoder could be realized using either analog techniques by means of bank of analog band-pass filters, or digital techniques by means of digital filters or FFT (Fast Foureir Transform) algorithms.

The main advantage of this technique is that the analysis is automatic. The drawback of the channel vocoder is that the speech produced sounds mechanical or monotonic.

1.2.1.2 Formant Synthesis Technique

This technique simulates the transfer function of the vocal tract. The transfer function of the vocal tract is characterized by several resonance frequencies, and it can be simplified by taking the first three formants (resonance frequencies of the vocal tract). The formant synthesis technique is closely related to speech production. As sound propagates from the glottis to the lips, the broad spectrum of the excitation source is shaped by the frequency selectivity of the vocal tract. The goal of formant synthesis is to generate a speech signal from the informations on the formant frequencies and bandwidths and sometimes ampli-

tudes [2]. The typical bit rate needed to control a formant synthesizer ranges from 1000 to 2400 bits/s [3].

Formant synthesizers are realized using adaptive analog filters, or digital filters by simulation on a computer.

The main advantage of this technique is its considerable degree of flexibility and efficiency in various applications of synthetic speech. It provides important insight into the basic mechanism of speech production and preception. The drawback of this technique is the difficulty of reliable automatic analysis of speech to detect the formants.

1.2.1.3 Articulatory Synthesis Technique

In this technique the vocal tract is represented by lossly acoustic tube of variable cross section area, having non-rigid walls. The propagation of sound waves inside this tube is simulated by solving the partial differential equations describing the physical phenomena covering this propagation. These equations can be solved using different approaches. The actual parameters used in this technique are the values of the air pressure and velocity functions corresponding to the message to be synthesized. These functions could be derived from an articulatory model using parameters such as: tongue position, jaw position, ... etc. These parameters change relatively slowly with time and could be coded by a bit rate as low as 50 bits/s [3]. Articulatory synthesizers can been realized employing analog or digital techniques.

1.2.1.4 LPC Synthesis Technique

Linear predictive coding (LPC) offers an efficient time domain technique for good quality and highly intelligible synthetic speech. [1]. The linear prediction model of speech synthesis was developed by Fant in the late 1950's [4]. Further development by Itakara and Satio (1968) [5], and then by Atal and Schroeder (1971) [6], resulted in modeling the speech waveform by a linear system with time varying parameters.

LPC is an analytic method of encoding speech as low bandwidth parameters, and of re-synthesising these with virtually no loss of information. LPC has become a standard method for encoding speech signals in parametric form, for both transmission and storage, at low bit rate [2]. The speech is produced (synthesised) using LPC technique as a result of exciting an all-pole digital filter by a sequence of impulses for voiced sounds and random noise for unvoiced sounds [7]. The typical bit rate needed to control an LPC synthesizer ranges from 1000 to 4800 bits/s [3]. LPC synthesizer is easily realizable using computer simulation or digital circuits.

The main advantage of this technique is its suitability to digital realization. Good speech quality can be obtained using 1000 bits/s or even 200 bits/s using vector quantaization of the synthesizer filter parameters [8]. Also automatic analysis of speech is possible which does not need formnat tracking or spectral fitting. The drawback of this technique is that the choice between either periodic or noise sources, i.e. voiced/unvoiced decision, is not always satisfactory [4].

1.2.2 Types of Speech Synthesis Methods

Speech produced by machine (synthesised) using the source/filter techniques can have varying degrees of speech data compression. Methods to synthesize speech based on source/filter techniques can be divided into two types [9].

1. Synthesis by analysis (analysis/synthesis).
2. Constructive synthesis.

The synthesis by analysis approach deals with long segments of speech such as words, phrases, or even sentences. The trade off here is between the intelligibility of speech and number of parameters associated with it.

The other approach, constructive synthesis method deals with the smallest units of sounds such as phonemes, allophones, diphone, etc. Systems constructed by this approach will generate unlimited vocabulary of synthetic speech with less storage requirements. The quality of synthetic speech using this approach is not as good as the one produced by the analysis/synthesis approach [9]. The main application of this method is text-to-speech systems.

1.3 Applications of Speech Synthesis

There are several applications of speech synthesis, some of them are

1. Flexible data base inquiry systems
2. Talking instrument panels
3. Talking computer terminals
4. Reading machines for the blind
5. Speaking aids for the handicapped
6. Remote reception of electronic mail by phone
7. Teaching machines and training aids
8. Talking books to teach reading

9. Remote access to information over the phone
10. Hobby computers

1.4 Properties of Arabic Speech

In many cases the development of a computer speech processing system depends on the natural language. For example, using modern techniques it is possible to estimate the shape of the vocal tract of a speaker from an analysis of the speech signals he or she produces [10,11]. Hence, systems developed for native speakers of one language may not be adaptable to the native speakers of another language. Some of the properties of Arabic speech have been reported in the literature [12,13]. In the following paragraphs we will present a summary of a study done by Penbeci and Hejres [12] for investigating properties of Arabic vowels.

The vowels are the most important class of articulatory gesture in any language. There are twelve vowels in Arabic, compared to ten in English. The Arabic vowels can be classified as an emphatic and non-emphatic, depending on the position of the tongue. The emphatic vowels have the following features

1. They are produced with the back of the tongue raised towards the roof of the mouth.
2. They have non-emphatic counterparts, from which they are clearly distinguished.
3. They produce more echo and thickness of voice than their no-emphatic counterparts.

The vowels are also classified, according to their duration, as short and long vowels and as kasra, fatha and dhamma.

The formant frequencies have been shown to be among the important features for many speech analysis and synthesis systems. In [12] the formant frequencies for the Arabic language vowels were studied. In particular the first three formants frequencies since they are usually enough to characterize a particular vowel. The spectrograms of a carefully chosen list of twelve Arabic words spoken by several male native speakers were obtained and analyzed. The words were chosen such that each contained one of the twelve vowels found in Arabic. A vowel triangle for Arabic language, shown in Figure 1.1 was plotted from the average locations of each Arabic vowels in formant space. Comparison of this vowel triangle with the one corresponding to the English language [1] which is shown in Figure 1.2, showes considerable differences. Although, in the formant plane, the upper left corner locations of the triangles are almost same, the other corners are completely at different locations.

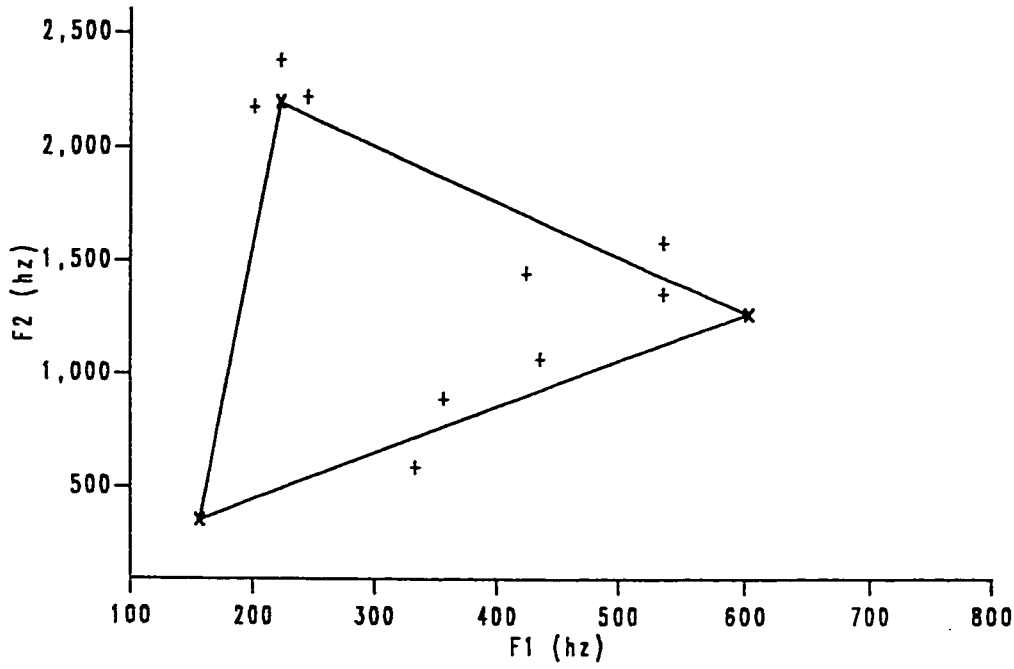


Figure 1.1: Arabic vowel triangle.

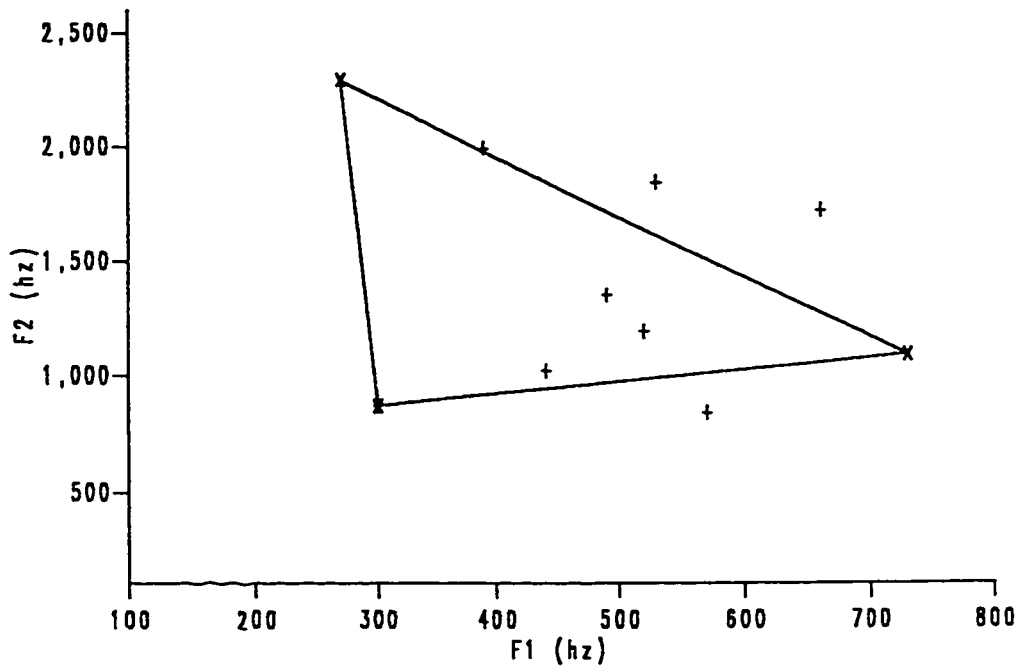


Figure 1.2: English vowel triangle.

1.5 Thesis Motivation

From the literature review, it appears that most of the research work on speech synthesis was done mainly for English language [1,2,4,6,14]. The necessity of extensive research in Arabic speech processing has been expressed in the literature [12,15-17]. However, only few articales appeared in the literature concerning Arabic speech synthesis [16]. Moreover, most of them concentrated on problems associated with constructive synthesis approach for text to speech synthesis [3,17].

In this thesis analysis and synthesis system based on LPC techniques is implemented and tested to produce good quality intelligible synthesized Arabic speech. This is achieved through finding suitable synthesis process which involves analysis of Arabic speech. The important parameters are: length of data frame used in the analysis mode; number of poles associated with the synthesis filter; pre-emphasis factor; kind of excitation function. These parameters for Arabic language are different from those obtained for English. Some of the reasons for this variation in analysis and synthesis system parameters are as follows: First, there are several sounds (voiced and unvoiced) in Arabic that do not exist in English. Also, the range of fundamental frequency (first formant) in Arabic vowels is larger than that of English vowels. Furthermore, similar vowels in both languages have different fundamental frequency values [12]. Therefore the evaluation of synthesizer's parameters are studied and tested for Arabic language.

Chapter II
SPEECH GENERATION

2.1 Introduction

In order to understand the LPC speech production model, we need to understand speech physiology, which is the basis of many different areas which are relevant to a better understanding of speech.

In this chapter the physiological and acoustical principles of speech will be considered briefly. Detailed discussions on speech are presented in Flanagan [18].

2.2 Human Vocal System

The human vocal system composed of two main parts, the vocal source and the vocal and nasal tracts (Figure 2.1). The vocal source includes essentially the lungs, the trachea and the larynx. The vocal and the nasal tracts form the supra-glottal cavities. The lungs play the role of air reservoir; air is forced from them, it passes through the trachea into the pharynx. The trachea is surmounted by the larynx. The larynx is a cartilaginous frame housing two lips ligament and muscles which are the vocal cords. The opening between the vocal cords is called the glottis. The supra-glottal tracts are composed of three principal parts, the pharynx, the oral cavity and the nasal tract. The nasal tract is coupled in parallel with the vocal tract by means of the velum.

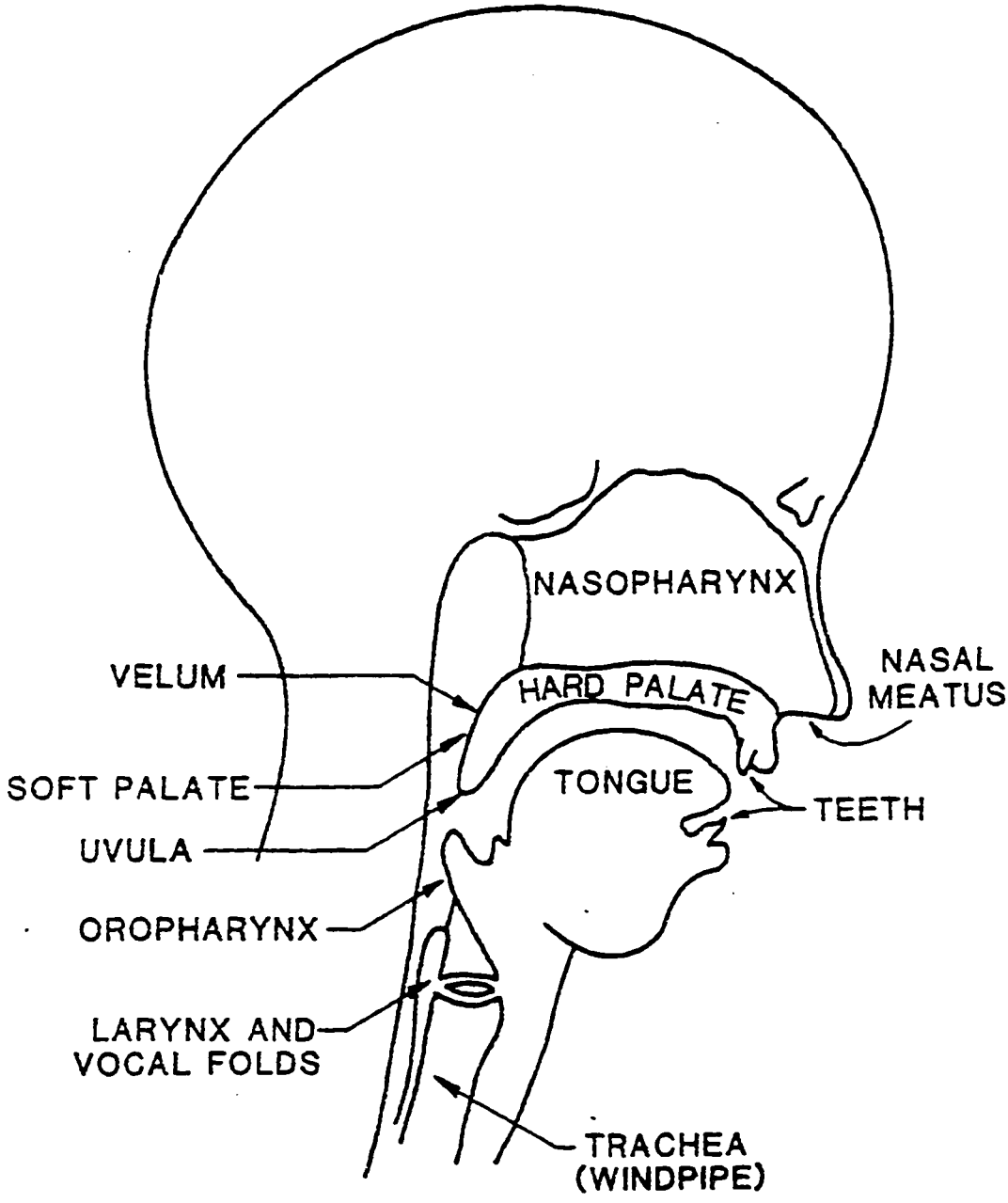


Figure 2.1: Human vocal system.

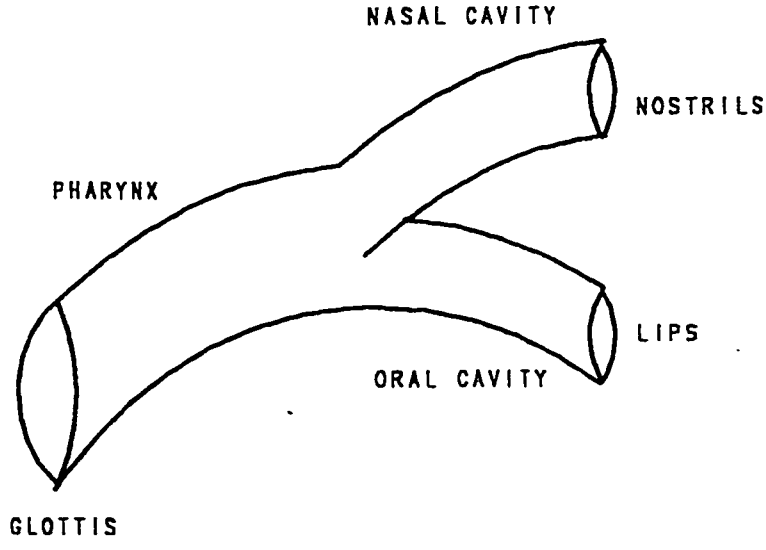


Figure 2.2: Nasal coupling to the vocal tract.

In general, the vocal system is an acoustical cavity (Figure 2.2). The vocal tract represents a non-uniform time varying (in shape) acoustical tube of an average length of 17 *cm*. The movements of lips, teeth, jaw, tongue, velum and larynx cause the change in the vocal tract volume. For example, the cross-sectional area of the lip opening can be varied over the considerable range from 0 cm^2 with the lips closed to about 20 cm^2 when the jaw and lips are open. These organ (lips,teeth ,etc.) are usually refered as the articulators. The nasal tract represents another acoustical tube of an average length of 12 *cm* that extends from the velum to the nasal meatus. The cross-sectional area of the nasal tract is variable over the first 3 *cm* which couples it to the vocal tract.

2.3 Speech Production Mechanism

The speech waveform is generated due to the variation of pressure above and below the vocal cords and due to the constriction produced within the vocal tract.. This constriction is a result of placing the articulators in certain ways. Sounds are classified into two major types according to the type of of excitation applyed from the vocal source.

2.3.1 Voiced sounds

When the vocal cords are tensed, their spacing is restricted and the flow of air causes the vocal cords to vibrate in such a way as to modulate the flow of air from the lungs. When the air flow is modulated in this way, the pressure variation of the flow of air into the vocal tract is quasi-periodic and the sound so produced is defined as a voiced sound as /a/ in apple. The frequency at which the vocal cords vibrate corresponds to the fundamental frequency and is related to the acoustic property identified as pitch of the sound. The vocal system being a resonant cavity, modifies the frequency content of the quasi-periodic waveform,

amplifying some frequencies and attenuating others. The vocal tract resonant frequencies are known as the formant frequencies, or in short formants. The articulators are used to shape the vocal tract which causes corresponding changes in the formants. These changes in the vocal tract shape produces different sounds like vowels, voiced fricatives and voiced plosives[1]. Another class of voiced sounds are nasal sounds. When the velum is closed, it decouples the vocal tract from the nasal cavity, nasal sounds and nasalized vowels are produced. The nasal sounds radiates through the nose. If the closure is at the lips the sound /m/ is produced and when it is behind the teeth the resulting sound is /n/.

2.3.2 Unvoiced Sounds

If the constriction caused by articulators is sufficiently narrow, turbulence results when the air from the lungs passes through this constriction, resulting in a noise-like excitation while the vocal cords are relaxed. Sounds produced such away are called unvoiced sounds. Also due to the vocal tract shaping, unvoiced sounds can be named as unvoiced fricatives and unvoiced plosives. An example of unvoiced fricative is /f/ as in five.

The conclusion from the above discussion of speech production is that changing the shape of the vocal tract and the function of the vocal cords produces different sounds. The velum aids in production of nasal sounds such as /m/ and /n/. Its upward and downward movement causes acoustical coupling and decoupling of the nasal cavity to the vocal tract. Thus the nasal sounds are radiates at nostrils instead of the mouth opening. In addition, sounds are recognized as voiced and unvoiced. Voiced sounds are the result of vibration of the

[1] When the air flow is completely restricted by closing the vocal tract so that the pressure is built behind the closure and the closure is suddenly opened to release the trapped air, the sound produced is called plosive, as /t/ and /b/.

vocal cords, whereas unvoiced sounds are the result of the constriction within the vocal tract, provided that the vocal cords are relaxed.

2.4 Speech Production Model

As it has been seen from the previous discussion that the speech waveform is the response of the vocal system when excited by one or two sources. It may be modeled as shown in Figure 2.3. The elastic structure of the vocal cords enables them to vibrate rapidly (50-500 hz) and produce, at the output of the larynx, a succession of air pulses. These pulses represent the first source which excites the supra-glottal cavities to produce voiced sounds. These pulses forms a periodic signal whose period $T_0 = 1/F_0$ where F_0 is the fundamental frequency or the pitch. The second source is that of noise, and is localized along the vocal tract, at a point which depends on the pronounced sound. The vocal tract can be represented, as mentioned before, by an acoustical tube with an interconnected sections of equal length and varying cross-sectional areas as a function of time. Figure 2.4 showes this representation of the acoustical tube.

2.5 Mathematical Model of The Vocal Tract System

The exact mathematical model of the vocal system is very complicated and would involve lot of details, which most likely would be neglected in favor of the simplicity and the low cost of the simulation networks [19]. Instead a simplified model is considered to capture the essential elements of the vocal system.

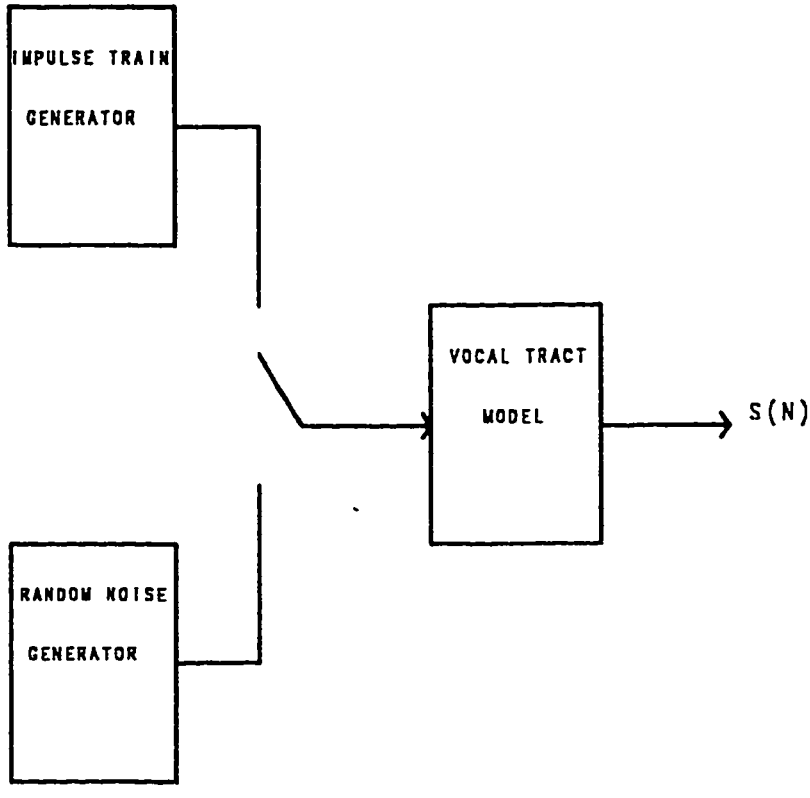


Figure 2.3: Speech production model.

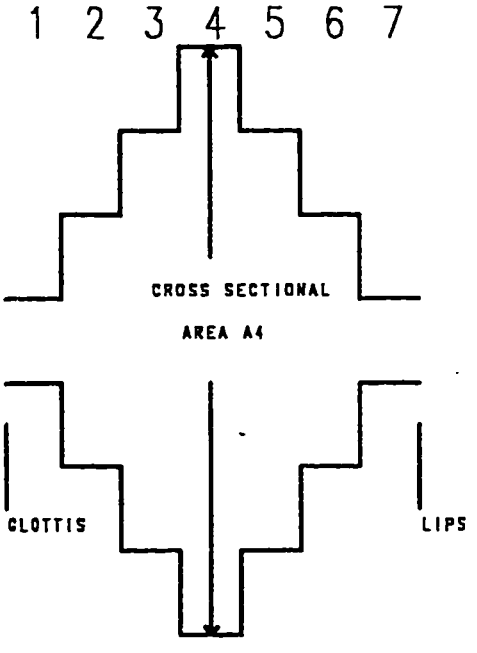


Figure 2.4: Concatenation of (N = 7) lossless tubes of equal length.

2.5.1 All-Pole Model of Vocal Tract

A simplified and practical model of speech mechanism will be considered here to simulate the main elements of this mechanism. It is shown in Figure 2.5 [4]. The assumption behind this model is given in detail by Flanagan [18]. It is clear from the block diagram that the digitized speech signal in z-domain is equal to

$$X(z) = E(z)\,G(z)\,V(z)\,L(z) \tag{2.1}$$

where $E(z)$ is the z-transform of the excitation source, $G(z)$ is the glottal and sub-glottal effects, $V(z)$ is the resonance filtering effect of the vocal tract, and $L(z)$ is the radiation factor of the lips. The vocal tract, the glottal and radiation effects can be lumped together to give

$$X(z) = E(z)\,H(z) \tag{2.2}$$

where

$$H(z) = G(z)\,V(z)\,L(z) \tag{2.3}$$

The glottal shaping model is approximated by

$$G(z) = \frac{1}{(1 - e^{-cT}z^{-1})^2} \tag{2.4}$$

Where both poles are real and inside the unit circle with the magnitude of the poles close to one.

The all-pole vocal tract model $V(z)$ consisting of k formants is described by

$$V(z) = \frac{1}{\prod_{i=1}^{k}[1 - 2e^{-c_iT}\cos b_iTz^{-1} + e^{-2c_iT}z^{-2}]} \tag{2.5}$$

Finally, the lip radiation $L(z)$ is of the form

$$L(z) = 1 - z^{-1} \tag{2.6}$$

The lip radiation accounts for the relation that occurs between the volume velocity of the air flow at the lips and the sound pressure of the radiated acoustical wave. To represent the time varying nature of speech, it can be assumed that the vocal tract is in steady state position over short periods of time.

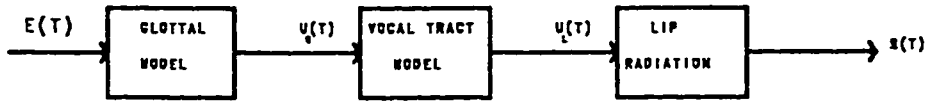


Figure 2.5: Block diagram for speech production mechanism

Thus, the system is modeled as a linear time-varying system whose parameters for short periods of time may be considered to be constant. Now the combined form of $H(z)$ is of the form

$$H(z) = \frac{1 - z^{-1}}{(1 - e^{-cT}z^{-1})^2\{\prod_{i=1}^{k}[1 - 2e^{-c_iT}\cos b_iT\, z^{-1} + e^{-2c_iT}z^{-2}]\}} \tag{2.7}$$

where the k formants are defined in the model. There is only one zero term in $H(z)$ and it is nearly canceled by one of the poles $1 - e^{-cT}z^{-1}$ since cT is generally much less than unity [4]. A further simplification can be made in an all-pole model as

$$X(z) = E(z) \cdot \frac{1}{A(z)} \tag{2.8}$$

by defining

$$A(z) = \frac{1}{H(z)} \tag{2.9}$$

The filter $A(z)$ is an all-zero filter and will be refered to as a inverse filter. The filter $H(z) = \frac{1}{A(z)}$ is an all-pole filter which represents the smooth spectral behavior of the speech model. In the next section we will consider the linear prediction technique which is equivalent to all-pole modeling (or autoregressive modeling) of speech mechanism considered here.

2.6 Linear Prediction of Speech

Linear prediction analysis applies to a class of problems in speech analysis and synthesis in which the present sample is predicted by a linear combination of past samples. The first research work to directly apply linear prediction technique to speech analysis and synthesis were by Satio and Itakura [5] and Atal and Schroeder [6].

In linear prediction a continuous signal $s(t)$ is sampled to produce a sequence $s(n)$. Then to provide an estimate of $s(n)$, $s(n)$ is chosen as the sum of past values of $s(n)$ and some unknown input as shown in equation (2.10). That is, the current signal value $s(n)$ is represented as a linear combination of some values of the output $s(n-k)$ and some weighted value of unknown input $u(n)$.

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + G\,u(n) \tag{2.10}$$

Where, $u(n)$ is some unknown input to the system, G is a gain factor and $\{a_k\}$ are an unknown coefficients.

Taking the z transform of equation (2.10) yields:

$$S(z) = -\sum_{k=1}^{p} a_k z^{-k} S(z) + G\,U(z) = H(z)\,U(z) \tag{2.11}$$

The above equation gives $H(z)$ as a representation of an all-pole transfer function.

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.12}$$

A block diagram illustration of the linear prediction of speech in time domain and frequency domain is given in Figures 2.6 and 2.7 respectively.

Given a particular signal $s(n)$ in equation (2.10), the problem is to determine the coefficients a_k and gain G to predict the next value of $s(n)$. However, due to the random nature of speech signal, the input is totally unknown. Therefore, $s(n)$ can be only estimated as a linear combination of past values.

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{2.13}$$

The error $e(n)$ between the actual $s(n)$ and the predicted $\hat{s}(n)$ is given as:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2.14}$$

The parameters a_k are to be obtained to minimize the total square error.

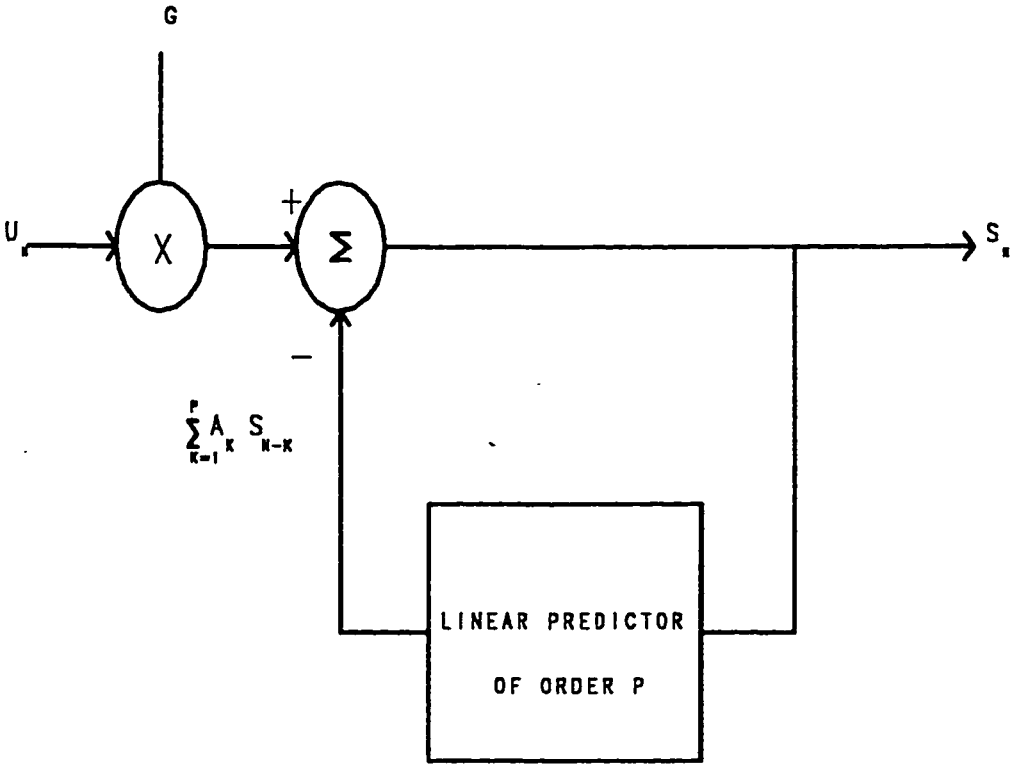


Figure 2.6: Linear prediction model in time domain.

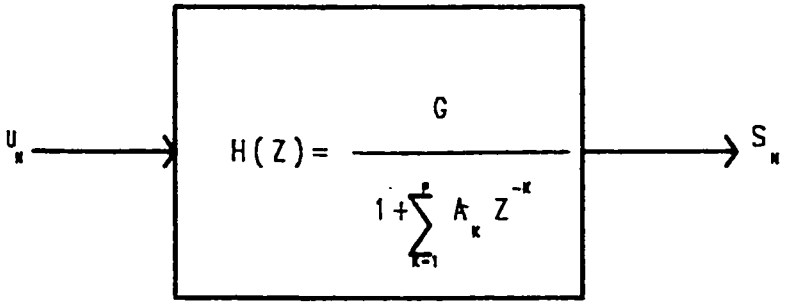


Figure 2.7: Linear prediction model in frequency domain.

$$E(e(n)) = \sum_n [e(n)]^2 = \sum_n [s(n) + \sum_{k=1}^{p} a_k s(n-k)]^2 \qquad -\infty \le n \le +\infty \tag{2.15}$$

The minimization of the total error is obtained by taking partial derivative of the total square error with respect to each a_k and equating it to zero.

$$\frac{dE}{da_i} = 0 \ , \quad 1 \le i \le p \tag{2.16}$$

which yields a set of equations.

$$\sum_{k=1}^{p} a_k \sum_n s(n-k)s(n-i) = -\sum_n s(n)s(n-i) \ , \qquad 1 \le i \le p \ , -\infty \le n \le +\infty \tag{2.17}$$

Equation (2.17) forms a set of "p" equations with "p" unknowns. The unknowns are the a_k coefficients called predictor coefficients of equation (2.13).

The minimum total square error is obtained by expanding equation (2.15) and substituting equation (2.17).

$$E_{\min} = \sum_n (s(n))^2 + \sum_{k=1}^{p} a_k \sum_n s(n)s(n-k) \ , \qquad -\infty \le n \le +\infty \tag{2.18}$$

Defining the autocorrelation function as

$$R(i) = \sum_n s(n)\, s(n-i) \qquad -\infty \le n \le +\infty \tag{2.19}$$

and

$$R(i-k) = \sum_n s(n-k)\, s(n-i) \qquad -\infty \le n + \infty \tag{2.20}$$

However,the signal $s(n)$ is known over a finite interval of time. A popular method is to multiply the signal $s(n)$ by a window function[2] so that the signal $s(n)$ will be zero outside some interval $1 \le n \le N$. The autocorrelation function is then given by

$$R(i) = \sum_{n=1}^{N} s(n)\, s(n-i) \ , \qquad i \ge 1 \tag{2.21}$$

---

[2] To be discussed later in Chapter III.

and

$$R(i\text{-}k) = \sum_{n=1}^{N} s(n\text{-}k)\, s(n\text{-}i), \quad i \geq 1 \tag{2.22}$$

Using equations (2.21) and (2.22) in equation (2.17) gives a short time auto-correlation equation in terms of the unknown coefficients a_k.

$$-R(i) = \sum_{k=1}^{p} a_k R(i\text{-}k), \quad 1 \leq i \leq p \tag{2.23}$$

Equation (2.23) is known as the autocorrelation method.

Writing equation (2.14) in the following form

$$s(n) = -\sum_{k=1}^{p} a_k s(n\text{-}k) + e(n) \qquad n = 1,2,3,\ldots,N \tag{2.24}$$

and comparing equation (2.24) with (2.10) we see that the unknown input signal is proportional to the error signal.

$$Gu(n) = e(n) \tag{2.25}$$

Whatever the total energy in the input signal $u(n)$ is, the energy of the output signal $\hat{s}(n)$ must be the same as that of $s(n)$ if the error is minimized.

Therefore, $Gu(n)$ is the total energy in the error signal as specified in equation (2.14). Applying a unit sample input to the system at $n=0$ i.e. $u(n) = \delta(n)$. The output of the linear predictor becomes:

$$h(n) = -\sum_{k=1}^{p} a_k h(n\text{-}k) + G\,\delta(n) \quad n = 1,2,3,\ldots,N \tag{2.26}$$

Where $h(n)$ is the unit sample response. By multiplying equation (2.26) by $h(n\text{-}i)$ and summing over all n, the autocorrelation equation, equation (2.23) becomes [7].

$$\hat{R}(i) = -\sum_{k=1}^{p} a_k \hat{R}(i\text{-}k) \quad 1 \leq i \leq p \tag{2.27}$$

Where $\hat{R}(i)$ is the autocorrelation of the predicted signal. From an impulse response, the total energy equation becomes:

$$\hat{R}(0) = -\sum_{k=1}^{p} a_k \hat{R}(k) + G^2 \tag{2.28}$$

Given that the total energy in $h(n)$ must be equal to that of $s(n)$ i.e. $R(0) = \hat{R}(0)$.

The squared gain is therefore given as:

$$G^2 = E_{min} = R(0) + \sum_{k=1}^{p} a_k R(k) \tag{2.29}$$

2.6.1 Linear Prediction Coefficient Evaluations

There have been several methods used to evaluate the predictor coefficients e.g. the Gauss reduction method and Crout reduction method which require lot of computations [7]. An efficient procedure was developed by Levinson known as the Levinson Durbin Recursion Method for solving M coupled sets of linear equations. [7]. This method makes use of the fact that the autocorrelation matrix in (2.27) is symmetric and the elements on the same diagonal are identical (i.e. a Toeplitz matrix). The LPC's of an inverse filter of order p are recursively obtained in p steps. The algorithm is executed by first applying an initialization:

$$E_0 = R(0) \tag{2.30}$$

$$k_1 = \frac{-R(1)}{R(0)} \tag{2.31}$$

$$a_{10} = 1 \tag{2.32}$$

$$a_{11} = k_1 \tag{2.33}$$

$$E_1 = E_0 (1-k_1^2) \tag{2.34}$$

where k_m are known as reflection coefficients and $R(i)$ are the autocorrelation values of equations (2.21) and (2.22). The technique then proceeds by forming the following values as the iteration progresses.

$$k_m = \frac{1}{E_{m-1}}[-\sum_{j=0}^{M-1} R(m-j)a_{m-1,j}] \tag{2.35}$$

The value p represents the number of equations present in the recursion process.

$$a_{m0} = 1 \tag{2.36}$$

$$a_{mj} = a_{m-1,j} + k_m c_{m-1,m-j} \quad j = 1,2,...,m-1$$

$$a_{mm} = k_m \tag{2.37}$$

$$E_m = E_{m-1}(1-k_m^2) \tag{2.38}$$

When the recursion is complete the values of the a_k coefficients in equation (2.27) becomes

$$a_k = a_{mk} \quad k = 0,1,2,...,M \tag{2.39}$$

From (2.38) one can write

$$E_m = R(0)\prod_{i=1}^{m}(1-k_i^2) \tag{2.40}$$

The minimum error E_m decreases monotonically as the predictor order increases. But for values of p larger than 14 the decrease becomes unsignificant [7]. For the filter described in equation (2.26) to be stable, the coefficients a_k must be such that $|a_k| < 1$, or equivalently

$$|k_m| < 1, \qquad 1 \le m \le p. \tag{2.41}$$

This stability condition can be shown to be a necessary and sufficient condition for the all-pole filter $H(z)$ to be stable, i.e. all poles are inside the unit circle. Filter stability is very important in speech synthesis, because unstable filter can lead to "pops" and "clicks" in the synthetic signal [7]. If $|k_p| = 1$, then the all p poles will be on the unit circle, which is unstable condition.

Due to the fact that linear prediction model and the acoustical tube model are equivalent, the reflection coefficients can be obtained from the area functions of the acoustical tube [1,4]. The relation is given by

$$k_i = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} \tag{2.42}$$

Where A_i is the ith cross-sectional area of the ith section of the tube. The order of the filter (or the number of reflection coefficient) corresponds to number of sections of the acoustical tube model. The linear prediction analysis is an efficient tool for estimating the equivalent area functions from the reflection coefficients.

LPC technique is superior, compared with the other techniques, in the sense that of providing a smooth all-pole approximation to the speech signal spectrum. Minimization of the prediction error ensures that the resulting model spectrum will fit the speech signal spectrum better at the signal peaks than the valleys [7]. This is a very desirable property since the spectral peaks for voiced sounds correspond to the formant frequencies of the vocal tract which are much more sensitive to the human ear than the valleys.

Chapter III
LPC SPEECH ANALYSIS AND SYNTHESIS SYSTEM

3.1 Introduction

A brief presentation of the speech analysis and synthesis was given in Chapter II. In this chapter an LPC analysis and synthesis system will be presented in some details.

The system is logically divided into two major modes.

1. Analysis mode (parameters extraction).
2. Synthesis mode.

The next section is devoted to the analysis mode and the successive section is for synthesis mode.

3.2 Speech Analysis

From the discussion covered previously by assuming that the speech is the result of exciting an all-pole LPC digital filter by an impulse train for voiced and random noise for unvoiced sounds, shows that we need to determine two important components for synthesising the speech, and they are:

1. Calculating LPC coefficients for the speech segments.
2. The voiced/unvoiced decision of the corresponding segments and pitch period estimation of voiced segments.

In the next sub-section we will consider the evaluation of LPC coefficient.

3.2.1 LPC Coefficients Evaluation

In fact there are two methods of solving for the prediction coefficients $a_k's$ known in the literature as the autocorrelation method and the covariance method [7]. In the covariance method, the prediction error E_p is minimized over a finite set of points. On the other hand, in the autocorrelation method the error is minimized over $-\infty$ to $+\infty$. The discussion made earlier considered in chapter two was limited to the autocorrelation method only since the stability of the all-pole filter is guaranteed, unlike the covariance method. Since only a finite set of samples is available, the infinite summation in the evaluation of the autocorrelation coefficients can not be preformed. Instead, the signal is windowed so that it becomes zero outside the range of available data [0,N-1] i.e. windowed signal will be

$$y_n = \begin{cases} w_n x_n & 0 \le n \le N-1 \\ 0 & otherwise \end{cases} \tag{3.1}$$

where w_n is a window function to be discussed later.

Proper analysis conditions for the linear prediction method are important to ensure satisfactory results. The analysis conditions to be noted are: sampling frequency, order of the all-pole filter, frame length, type of windowing and preemphasis.

3.2.1.1 Sampling Frequency

The sampling frequency determines the frequency range of interest. As the sampling rate increases, the representation of a continuous speech signal becomes more accurate. However, more samples imply larger storage requirements and greater computation. For accurate estimation of voiced speech, the sampling frequency should exceed 6 khz to include a speech bandwidth of at least 3 khz . A general rule of thumb is to sample as slowly as possible without destroying

the significant features of the signal [4]. A sampling frequency of 6.5 khz is used in our study.

3.2.1.2 Order of the Filter

The order of the all-pole filter is dependent on the frequency range to be chosen. When the frequency range is exactly half the sampling frequency (F_s), in khz a good rule of thumb for the number of poles is from $F_s + 2$ to $F_s + 4$ [4]. The reason for this appears to be that there will be about $F_s/2$ resonances in the frequency band limited by $F_s/2$. Each resonance requires two poles for its representation, and so about F_s poles will be needed to account for the expected resonances in the analysis band. In addition, two to four coefficients are normally used for approximating the spectral slope due to the excitation source. In our study we tested different number of poles that suite Arabic speech synthesis.

3.2.1.3 Frame Length

For the purpose of calculating the LPC coefficients and the pitch period, the discrete speech signal is blocked into NF units, called frames, where each unit is of length L points. The speech signal is non-stationary process and the autocorrelation method cannot be performed reliably unless the speech signal satisfies stationarity condition. However, speech signal can be considered stationary for short periods of time [7], so the frame length is chosen such that the stationarity of the speech signal is justified and such that the estimation of autocorrelation coefficients can be reliably preformed. The autocorrelation method requires a frame length of at least 1-5 pitch periods for voiced sounds. Thus analysis is performed on the data blocks of length LF every $1/F_r$ units of time where F_r is the analysis frame rate. Frame rate is used to have an overlap between two consecutive frames for producing smooth transition in the LPC

parameters values, but this will require more computations and storage requirement. A simpler way to smooth the values of LPC parameters while keeping minimum computations and storage requirements is to have no overlap between the analyzed frames but apply a linear interpolation between the LPC parameters in the synthesis process [5]. The frame length is tested in this study with having no overlap between consecutive frames. [23]. k indicates the frame number.

3.2.1.4 Windowing

The framing operation introduces an implicit rectangular windowing whose effect in the time domain is to create discontinuities at the boundary points of the frame. In the autocorrelation method of linear prediction, where the signal is assumed to extend from $-\infty$ to $+\infty$, the abrupt changes at the boundary points create a spectral distriction in the high frequency region. Windowing purpose is to taper the data at the end points and caused the required smooth transition. Different types of windows were proposed and their spectral characteristics were also discussed in the literature [1,4] i.e. rectangular, Hamming, Hanning and etc. The Hamming window was shown to produce a smoother spectrum than the other windows. Therefore, it was used in this thesis and it is given by [23]

$$w_n = 0.54\text{-}0.46\cos(2\pi\frac{n}{L\text{-}1}),\quad 0\le n\le L\text{-}1 \tag{3.2}$$

to give the windowed signal

$$y_n = \begin{cases} w_n x_n & 0\le n\le L\text{-}1 \\ 0 & otherwise \end{cases} \tag{3.3}$$

The time domain plot of Hamming window is given in Figure 3.1.

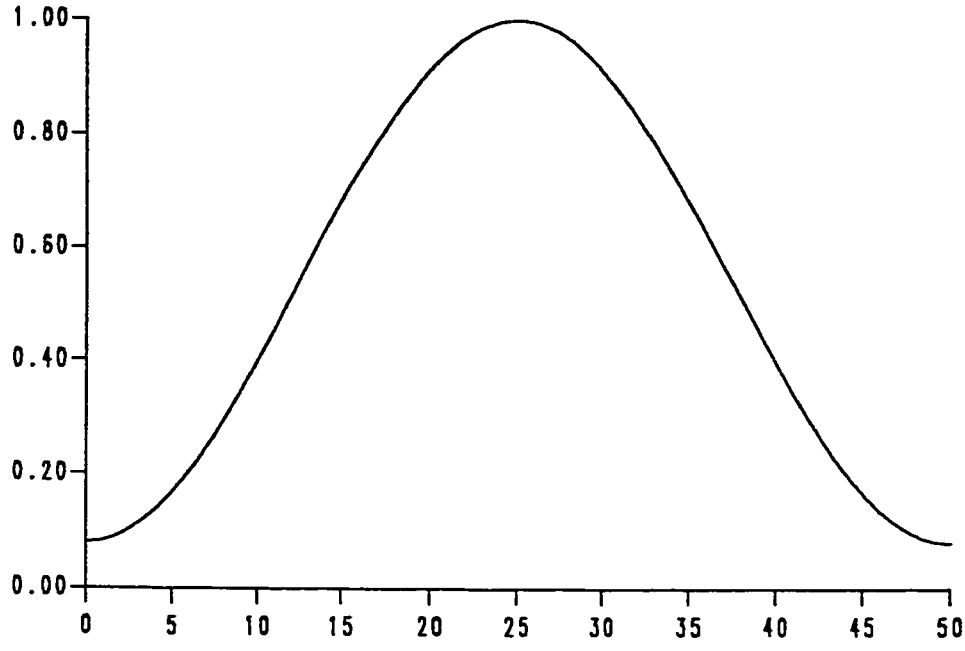


Figure 3.1: Hamming window in time domain.

3.2.1.5 Preemphasis

For the purpose of enhancing the spectral peaks in the higher frequency region, a 6 dB/octave emphasis filter of the form $1-az^{-1}$ is used. From the literature a can take the values $0.9 \le a \le 1$ and yield roughly equivalent results [4].

From the linear model of speech production described in chapter two equation (2.1), it was shown that the combined effect of the vocal tract the glottal wave and the radiation factor are given by

$$H(z) = V(z)\, G(z)\, L(z) \tag{3.4}$$

where $G(z)$ is a two-pole function whose poles have magnitude near one, $V(z)$ is the all-pole model of the vocal tract and $L(z)$ has one zero. The zero of the preemphasis filter cancels one of the two glottal wave poles. The other pole is compensated by the zero due to lip radiation factor. It is clear that this comment concerns voiced speech, since in unvoiced speech we do not have the glottal wave effect. Markel and Gray [14] have considered an optimal or adaptive pre-emphasis factor given by $\mu = R_y(1)/R_y(0)$ where $\{R_y(n)\}$ is the autocorrelation sequence of the data $\{y(n)\}$. For unvoiced sounds, μ will be small, whereas for voiced sounds μ will be near unity [4].

3.2.2 Pitch Detection and Voiced/Unvoiced Decision

The fundamental frequency (F_0) (or pitch frequency) is a basic parameter in acoustical studies of speech. It is also a necessary parameter for speech synthesis systems. It is the acoustical correlate to the rate at which the vocal cords vibrate. If the cords are vibrating rapidally, a high fundamental frequency will be measured. The reciprocal of the fundamental frequency is the pitch period P. The pitch period is a measurement of the time interval between successive complete cycles of cords opening and closing.

In the linear speech prediction model, the fundamental frequency is the rate at which the glottal volume velocity pulses are applied to the vocal tract, i.e. the driving function to the model is periodic with period of $1/F_0$. The ear is by order of magnitude more sensitive to changes of fundamental frequency than changes of the other speech parameters [18]. The quality of speech synthesis system is essentially influenced by the quality and faultless-ness of the pitch measurements.

Because of the importance of pitch estimation, a wide variety of algorithms for pitch detection has been proposed in the speech processing literature [24]. In these algorithms voiced/unvoiced decision is done in combination with pitch period estimation. One of the few studies which evaluates the performance of pitch detection algorithms in detail was done by Rabiner et al [25] and McGonegal et al [26]. A reference pitch contour was obtained interactively using the semi-automatic pitch detection algorithm by McGonegal et al [27]. Seven pitch detection algorithms (PDAs) were involved in that investigation. The main result read as follows, non of the PDAs involved worked without errors, even under good recording conditions. Each algorithm had its own error. So we conclude from this that there is no single algorithm that it is better than the others. We have chosen three pitch detection algorithm and tested them for Arabic speech. Those algorithms are, the simplified inverse filtering technique (SIFT) algorithm [28], the average magnitude difference function (AMDF) algorithm [29] and the maximum likelihood (ML) algorithm [30]. The performance of the ML and AMDF algorithms gave much better results for our system than the SIFT algorithm. This is because both the ML and AMDF algorithms are designed for realistic situation [31] which is the case in our work. The realistic situation means; noisy environment, absence of the fundamental frequency

due to band-limiting, simultaneous presence of periodic and random excitation, phase distortion, or rapid changes in pitch period. In particular noise is a serious limitation in many applications of analysis-synthesis systems [30]. In the following paragraphs we will briefly present the ML and AMDF algorithms.

3.2.3 ML Algorithm for Pitch Period Detection

Let s_k be a periodic repetition of sequence q_k, and length T i.e. $s_k = q_{k \bmod T}$. Then received signal r_k of length N,

$$r_k = s_k + n_k \quad k=0,...,N-1 \tag{3.5}$$

where N is the frame length and n_k are the noise samples which are independent identically distributed Gaussian random variables with zero mean and σ^2 variance. From the received signal it is desired to estimate the pitch period T. The period T and be estimated [30] by maximizing the function

$$g(T) = \frac{2T}{N} \sum_{i=1}^{I-1} Q_x(i+T) \tag{3.6}$$

where[3]

$$I = [N/T] \tag{3.7}$$

and

$$Q_x(k) = \sum_{j=0}^{N-1-k} r_j r_{j+k} \tag{3.8}$$

Thus the pitch estimation rule is based on computing g(T) for different values of T and the pitch period $\hat{T}$ is the value at which g(T) is maximized. The algorithm is not provided with voiced/unvoiced decision, so we investigate our own decision depending on our observations of the values of g(T). This algorithm has the capablity of determining the pitch period with a resolution finer than one sample.

---

[3] [.] is the greatest integer

one sample.

3.2.4 AMDF Algorithm for Pitch Detection

The AMDF is a variation of autocorrelation analysis. Instead of correlating the input speech at various delays (where multiplications and summations are formed at each value of delay), a difference signal is formed between the delayed speech and the original and at each delay the absolute magnitude of the difference is taken and it is defined by the relation

$$D_\tau = \frac{1}{L}\sum_{j=1}^{L}|s_j - s_{j-\tau}|, \quad \tau = 0,1,..,\tau_{max} \tag{3.9}$$

Where $s_{j-\tau}$ are the samples time shifted τ seconds. The difference signal exhibits deep nulls at delays corresponding to the pitch period of voiced sounds. AMDF is attractive because it is a simple measurement which gives a good estimate of pitch period and it has no multiplication operations.

3.3 Speech Synthesis

We accomplish the linear prediction synthesis scheme of speech as shown in Figure 3.2. From the extracted parameters in the analysis mode, i.e. prediction coefficients, voiced unvoiced decision and pitch period, and using a specific excitation input speech waveform is obtained. The stability of synthesis filter is guaranteed in the case of autocorrelation method used in this study.

3.3.1 Synthesis Filter Structure

In the digital signal processing literature, there are numerous filter structures that can be used to implement the general linear transfer function of the form $P(z)/A(z)$ [4]. Research in linear prediction of speech has uncovered a number of new structures for implementing $P(z)/A(z)$. These structures are of considerable importance in speech synthesis for two reasons

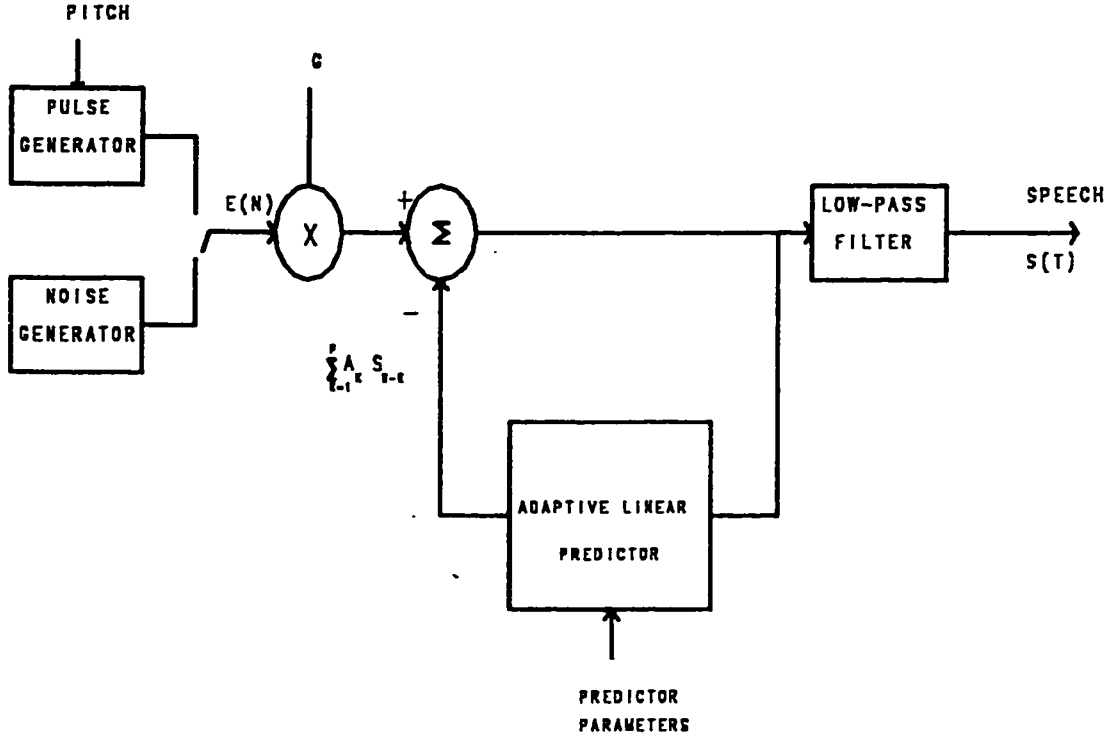


Figure 3.2: Speech synthesis scheme.

1. They allow the filter to be implemented directly from the reflection coefficients $\{k_m\}$.
2. They allow one to trade accuracy and complexity in an actual computer implementation.

The synthesis filter should mathematically implement the reciprocal of the inverse filter $A(z)$. Implementation of $1/A(z)$ in this study was done using the reflection coefficients $k_i's$ instead of the prediction coefficients $a_i's$, since $k_i's$ are less sensitive to noise (rounding off noise in this case) which affects the filter stability. A two-multiplier lattice model shown in Figure 3.3 is used in our study.

3.3.2 Excitation and Synthesis Matching

In the correlation matching formulation [4] a match between the autocorrelation of the input sequence $\{x(n)\}$ and the unit sample response of the all-pole synthesis filter $G/A(z)$ is desired at as many points as possible.

Each synthesised speech sample has two major components:

1. The decaying complex exponential value $\{q(n)\}$ from the previously synthesised pitch period
2. The synthesiser output $\{u(n)\}$ in response to an excitation sequence $\{e(n)\}$, without any effects from the previous frame.

Here the excitation source is either a series of periodic unit samples followed by zeros for voiced sounds or a series of samples from a random number generator for unvoiced sounds. Introducing a gain factor G, the total synthesizer response $\{\hat{s}(n)\}$ for the new frame is then given by:

$$\hat{s}(n) = q(n) + Gu(n) \tag{3.10}$$

Using an overbar "—" to denote the sum over N samples, e.g.,

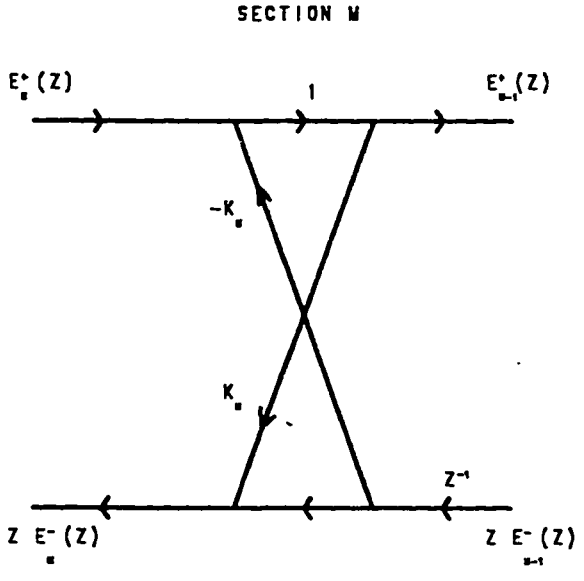


Figure 3.3: A two-multiplier lattice filter.

$$\bar{u}(n) = \sum_{n=0}^{N-1} u(n) \tag{3.11}$$

equal speech and synthesis energy requires

$$\overline{s^2(n)} = \overline{\hat{s}^2(n)} = \overline{[q(n) + Gu(n)]^2}$$

$$= G^2\overline{u^2(n)} + 2G\,\overline{q(n)u(n)} + \overline{q^2(n)} \tag{3.12}$$

Atal and Hanaur [6] proposed a solution for this quadratic equation solving for G. Their approach requires lot of computations and it is limited to a specific synthesis filter structure [4]. Markel and Gray proposed a simpler but less accurate method to drive the synthesis filter with the input sequence $\{e(n)\}$ to compute u(n), where u(n) now contains both the response due to the previous frame and the present excitation. Applying the same energy matching criterion

$$\overline{s^2(n)} = \overline{\hat{s}^2(n)} = G^2\overline{u^2(n)} \tag{3.13}$$

The term G is then computed as

$$G = \left|\frac{\overline{s^2(n)}}{\overline{u^2(n)}}\right|^{1/2} \tag{3.14}$$

The above approaches directly match the energy on an input-output basis. Based upon the analysis-synthesis model, the inverse filter output with energy G^2 drives the reciprocal inverse filter or synthesis filter to generate the speech output. A computationally efficient scheme for computing the deriving function with the autocorrelation method based upon the gain G was proposed by Markel and Gray. Assuming N data samples per analysis frame, a total square error of G^2 is obtained. If the random number generator samples $\{g(n)\}$ used for unvoiced excitation have a variance σ_g^2, then to match the energies over N samples, the driving function $\{e(n)\}$ is computed by

$$Ne^2(n) = g^2(n)\frac{G^2}{\sigma_g^2}$$

or

$$e(n) = g(n)\frac{G}{\sigma_g N^{1/2}} \tag{3.15}$$

To compute $e(n)$ for a voiced frame the inverse filter output is modeled as a train of unit samples separated by an integer number of samples I corresponding to the pitch period. Then, e(n) on the average is computed by

$$e^2(n)\delta_{n,mI}N/I = G^2$$

where m = 0,1,2,...,or

$$e(n) = \begin{cases} G(I/N)^{1/2} & n=0,I,2I,.. \\ 0 & elsewhere \end{cases} \tag{3.16}$$

3.3.3 Post-emphasis

If pre-emphasis has been applied in the analysis mode, post-emphasis must be applied at the output of the synthesis filter. The post-emphasis is performed by applying the reciprocal of the pre-emphasis filter as

$$\frac{1}{P(z)} = \frac{1}{1-az^{-1}} \tag{3.17}$$

Chapter IV
IMPLEMENTATION AND TESTING OF THE SYSTEM

4.1 Introduction

This chapter is devoted for the description of the experimental implementation of the LPC analysis and synthesis system and testing its parameters for good quality Arabic speech.

4.2 Data Acquisition

4.2.1 Analog Data Collection

Five statements were carefully chosen to cover most (if not all) Arabic sounds [32]. These statements were spoken by three male native Arabic speakers in standard Arabic and they are:

1. Addadu lukhat al-arab الضاد لغة العرب
2. kharaja tariq خرج طارق
3. Hatha zytun mobarak هذا زيت مبارك
4. Asharea nazeefun الشارع نظيف
5. Aussahra shasea الصحراء شاسعة

These statements were input to COMPAC deskpro 386 PC directly from the speakers, during digitizing process, in a normal laboratory environment.

4.2.2 Digitization and Data Transfer

The collected analog statements were low-pass filtered with cutoff frequency of 3.25 khz. The filtered signal was then digitized at 6.5 khz using a Data Translation A/D converter installed on a COMPAQ PC at 12 bits/sample. This convertion was controlled by using a special software package developed specially for signal processing purposes known as ILS -PCI package [33].

Because the analysis and synthesis system was implemented on the IBM 3033, the data was transfered from the COMPAQ PC to the mainframe using the IRMA interface card between the two. A block diagram in Figure 4.1 shows the representation of the data acquisition setup.

4.3 Software Implementation of The System

Both of the analysis and synthesis modes were implemented as FORTRAN programs on the mainframe. In the analysis mode, the speech samples were divided into NF frames each frame of LF samples in length with an overlap of LF/2. Each analysis frame is then pre-emphasized according to Markel and Gray criterion discussed in section 3.2.1.5. Then the pre-emphasized frame is windowed by Hamming window for evaluating the LPC coefficients. The voiced/unvoiced decision and the pitch period estimation is carried out by using AMDF algorithm since its computation time is much less than the ML algorithm. Figure 4.2 shows the block diagram of the analysis mode.

The synthesis mode implementation is illustrated by the block diagram, for the basic synthesizer operation (ignoring the setup of initial conditions), shown in Figure 4.3.

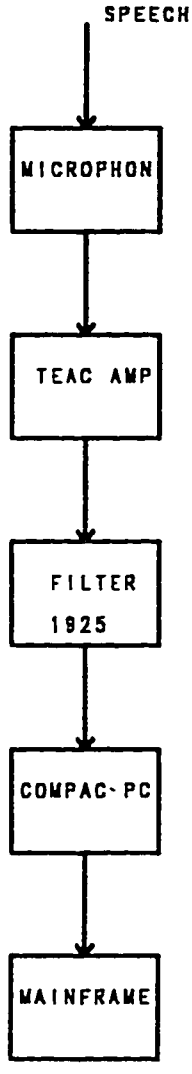


Figure 4.1: Data acquisition setup

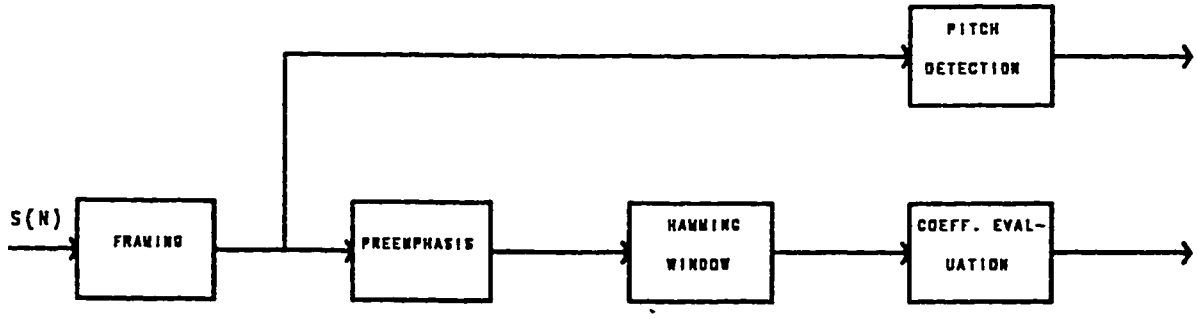


Figure 4.2: Block diagram of the analysis mode

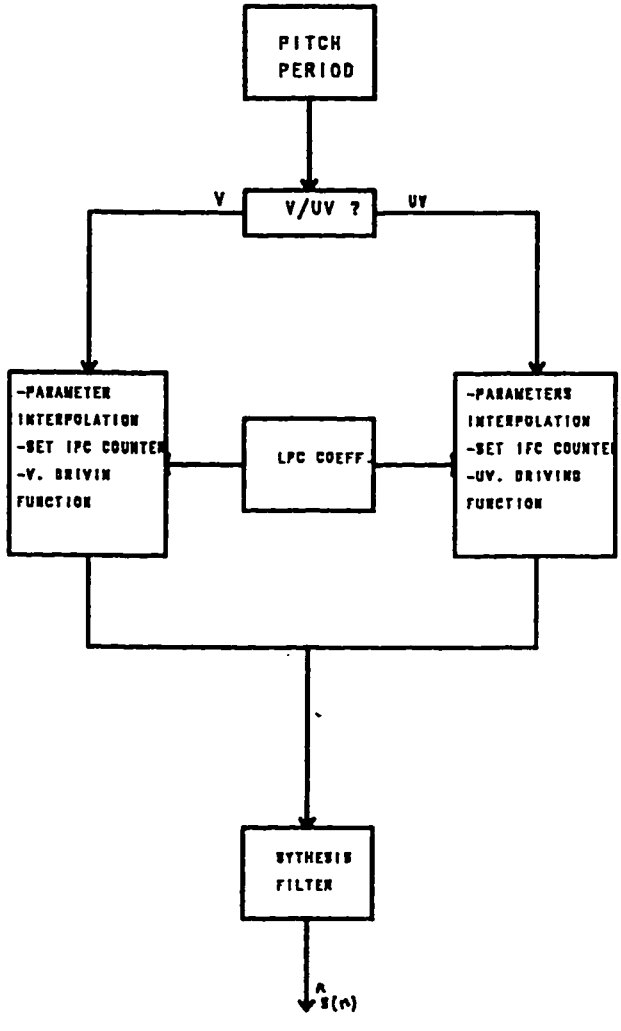


Figure 4.3: Block diagram of the synthesis mode

Linear interpolation is used in synthesis process for the analysis parameters. Interpolation requires knowledge of two frames of stored parameters. These frames are refered to as left-hand and right hand frames. Synthesis is applied only for the left hand parameters. Each time new parameters read by the right-hand frame the previous parameters, in the right-hand frame from the last step, are shifted into the left-hand frame. Then interpolation is applied and synthesis process takes place for the interpolated parameters in the left-hand frame. If the current frame, to be synthesized, is unvoiced a special counter for unvoiced frames IFC will start counting until the current frame is ended. If the current frame in the synthesis process is voiced another counter IPC will start counting from the start of each pitch period to its end. The program checks if the last frame of parameters is used for synthesis or not, if yes the process will terminate otherwise it will continue. The counters IFC and IPC are to control the start and the end of the current synthesized frame and the pitch period respectively.

4.4 Testing The System Parameters

This section is devoted to present the testing of both analysis and synthesis parameters for good quality Arabic speech. Initially the system developed by Markel and Gray [14] was considered which has the following parameters values for synthesizing good quality English speech.

1. Sampling rate of 6.5 khz.
2. Analysis frame length of 128 samples.
3. Frame rate of 51 hz.
4. Number of poles equals 10.
5. Pre-emphasis factor of 0.9.

6. Driving function for the synthesizer is periodic impulses for voiced frames and pseudo-random numbers uniformly distributed for unvoiced frames.

The implemented system was set[4] according to these parameters but the results, as we expected, were not satisfactory. Accordingly the system was remodeled to produce good quality Arabic speech. To decide on the suitable value of a certain parameter, we depended on two types of tests: a subjective test and a spectrographic test. The subjective test is accomplished by developing a questionnaire for a group of 16 listeners and study their judgements to form a statistical evaluation about the suitable value of the parameters. For each tested parameter three different statements from the three different speakers were synthesized for different values of that parameter. It was asked, in the questionnaire, to order the synthesized statements for each speaker according to intelligibility and quality starting with the best ones. Finally the satisfactory output values of the parameters from the questionnaire were used to produce the final adequate synthesized statements. Figure 4.4 shows the used questionnaire.

Spectrographs are indispensable aids in analysis both real and synthetic speech [2]. Usually, when a spectrograph of synthetic speech resembles that of real speech,the synthetic and real speech will sound the same. Important features to be noted in the spectrographs are the distribution of the speech energy different frequency bands (indicated by dark regions) and the formant transition with time.

---

[4] The implemented system uses AMDF pitch detection algorithm instead of SIFT algorithm used in [14].

Questionnaire

You will pass through three tests. In each test there are three groups of different statements. Each group contains several utterances of one statement. You are asked to order these utterances according to good quality and intelligibility starting with the best.
If you think that two statements or more have the same quality and intelligibility, put them in one horizontal line.

Test no.1

group1	group2	group3
quality and intelligibility	quality and intelligibility	quality and intelligibility

Test no.2

group1	group2	group3
quality and intelligibility	quality and intelligibility	quality and intelligibility

Test no.3

group1	group2	group3
quality and intelligibility	quality and intelligibility	quality and intelligibility

Figure 4.4: Questionnaire

The spectrographic test is done by taking the spectrograph of each synthesized statements for each tested parameter and compare them with the spectrograph of the original spoken statements to find the one which is similar or closer to the original spectrograph.

In the next section we will consider the testing of parameters procedure. The analysis parameters were considered first then the synthesis parameters.

4.4.1 Analysis Parameters

In the analysis process, the sampling frequency was kept at 6.5 khz. The reason is that higher sampling rate means more computations and more storage requirements. Thus this sampling frequency is suitable to get the most important features of speech while reducing the computation and storage requirements. Besides, it will help in making a comparison between the reported system conditions [14] and conditions of our system. Also, the preemphasis factor is kept the same as in the reported system because values of the preemphasis factor in the range between 0.9 to 1 produces roughly equivalent results [4]. The following analysis parameters were found have important effects on the quality of the synthesized speech.

4.4.1.1 Frame Length

The analysis frame length of 128 samples used by Markel and Gray [14] produced unacceptable Arabic synthesized speech. According to this, we tested several values of frame length to find the suitable value for our system. We synthesized three statements for different speakers using analysis frame length values of 100, 128, 150 and 175 samples. The corresponding values of the frame rates to the values of frame lengths for no overlap between the consecutive frames are:

frame length	100	128	150	175	samples
frame rate	65	51	43	37	hertz

A questionnaire test is used to find the best value of frame length. From the results shown in Table 4.1 it is clear that frame length of 150 samples gave the best result. The corresponding frame rate value is 43 hz. For example, in the first group results 87.5 % of the listeners had chosen the synthesized statement using frame length of 150 samples to be better than the other synthesized statements. According to this it was replaced in the second column indicating that it has order two after the original speech which has order one. For the second group the corresponding synthesized statement to frame length of 150 samples used in the analysis mode was chosen by 93.75% of the listeners to have good quality and intelligibility than the other synthesized statements, so it was given lower order (second best) than the others. The same thing can be said to the third group. Also, the spectrographic test for synthesized statements confirmed these results. One set of the spectrographic test is shown in Figure 4.5.

4.4.1.2 Number of Poles

The number of poles of the analysis filter is tested in the same manner as it was done for testing the frame length. Synthesized statements were produced using number of poles of 6, 8, 10, 12 and 14. Table 4.2 shows the obtained results from the questionnaire. Although the obtained results were close, eight analysis poles gave the best results. The spectrographic test confirmed this decision. Figure 4.6 shows one set of the spectrographic test.

Table 4.1: Frame length testing results

group1

frame length	order				
	1	2	3	4	5
original	100%	-	-	-	-
100	-	-	-	50%	50%
128	-	-	31.25%	56.25%	12.5%
150	-	87.5%	12.5%	-	-
175	-	12.5%	75%	12.5	-

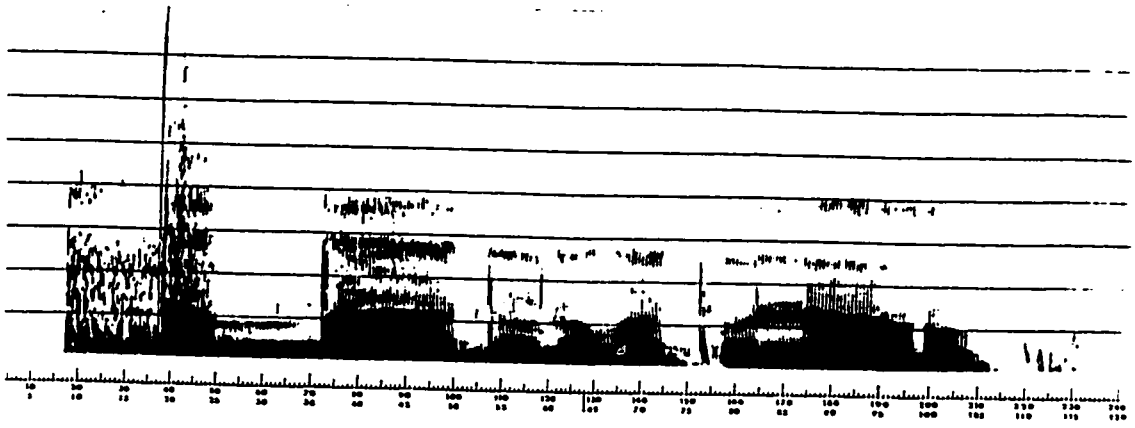
group2

frame length	order				
	1	2	3	4	5
original	100%	-	-	-	-
100	-	-	12.5	31.25%	43.75%
128	-	6.25%	12.5	75%	6.25%
150	-	93.75%	6.25%	-	-
175	-	18.75%	81.25%	-	-

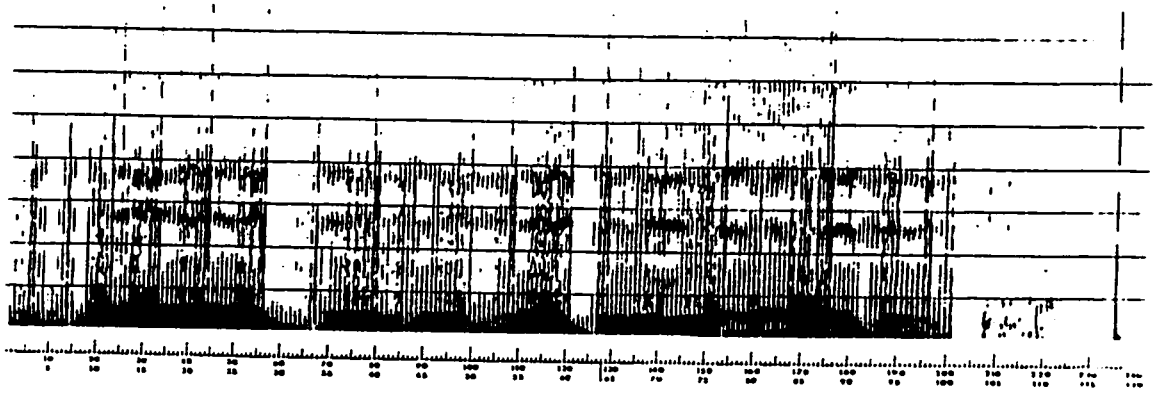
group3

frame length	order				
	1	2	3	4	5
original	75%	18.75%	6.25%	-	-
100	12.5%	-	18.75%	25%	43.75%
128	-	43.75%	43.75%	6.25%	6.25
150	12.5%	56.25%	18.75%	12.5%	-
175	12.5%	18.75%	31.25%	37.5%	-

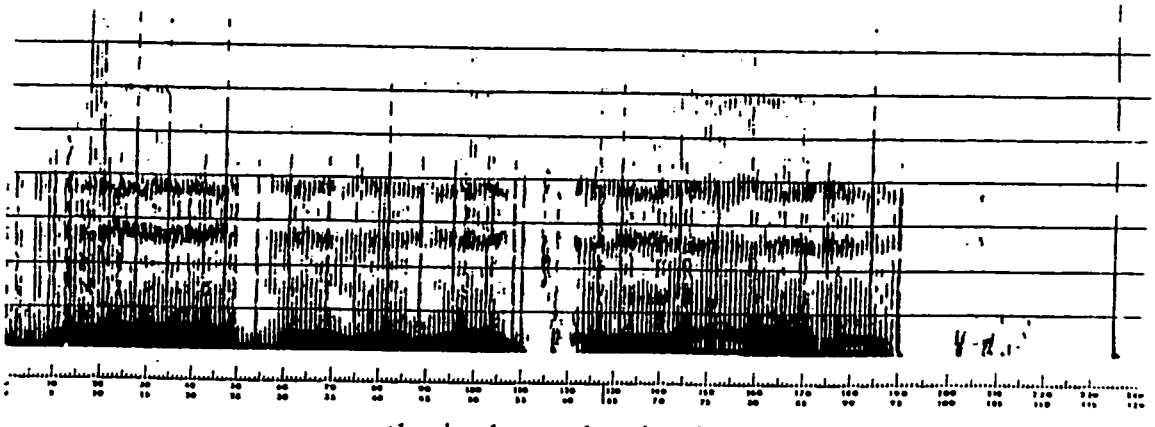
الضاد لغة العرب



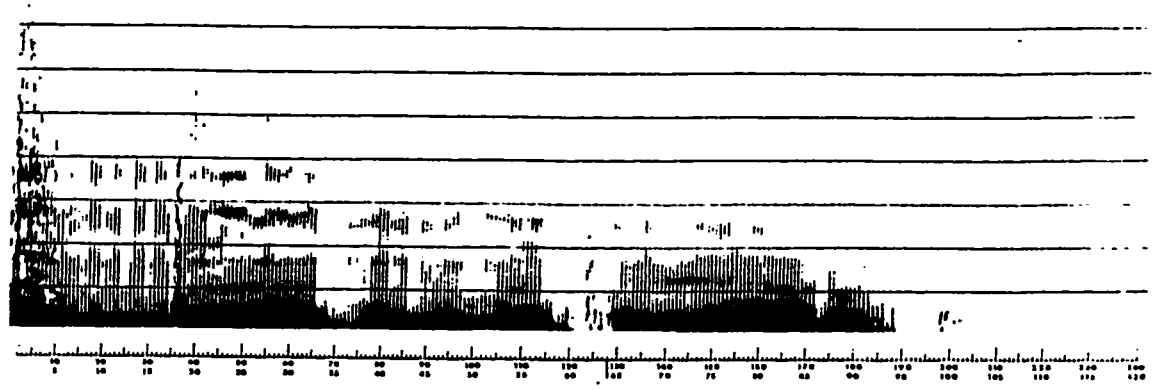
original



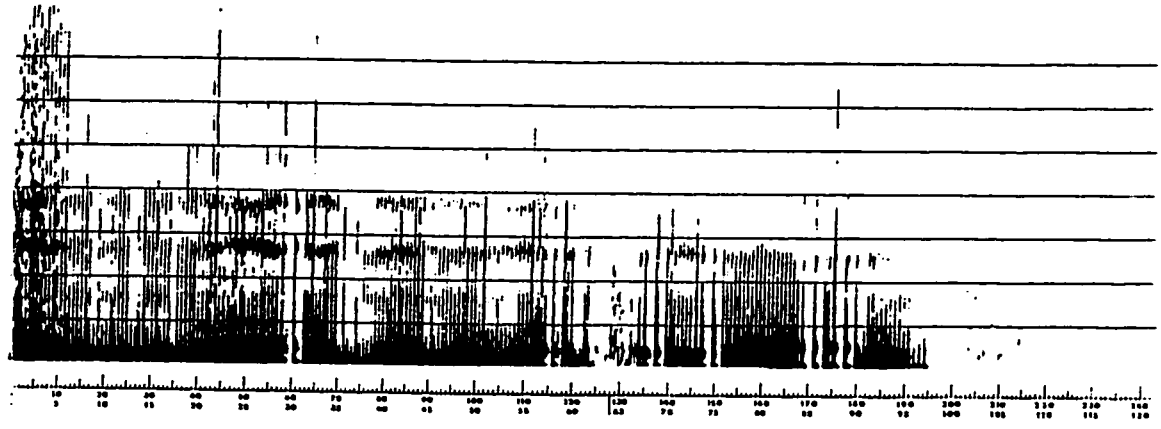
synthesized speech using LF = 100



synthesized speech using LF = 128



synthesized speech using LF = 150



synthesized speech using LF = 175

Figure 4.5: Spectrographs of one sample set of the first test.

Table 4.2: Number of poles testing results

group1

poles number	order				
	1	2	3	4	5
6	25%	43.75%	6.25%	12.5%	6.25%
8	68.75%	31.25%	-	-	-
10	68.75%	12.5%	12.5%	6.25%	-
12	43.75%	18.75%	31.25%	6.25%	-
14	43.75%	-	43.75%	12.5	-

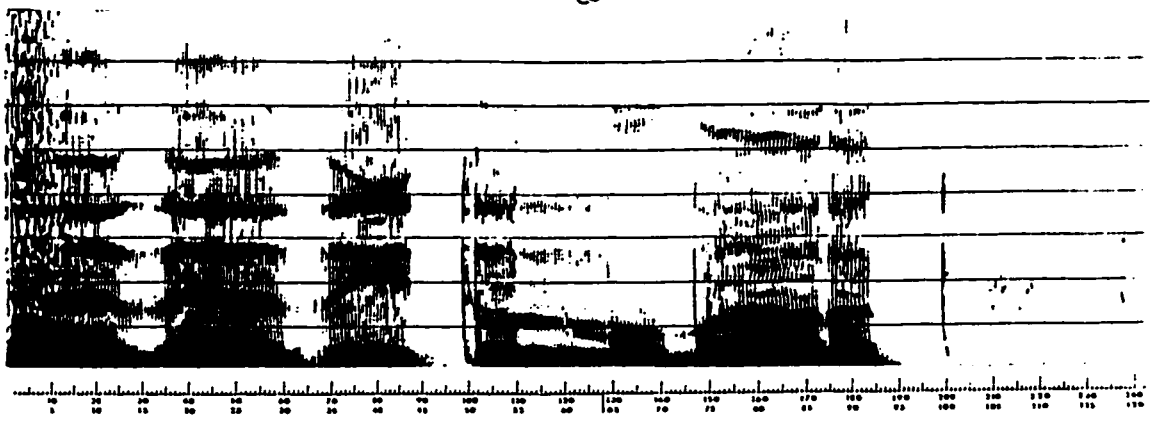
group2

poles number	order				
	1	2	3	4	5
6	18.75%	18.75%	43.75%	12.5%	6.25%
8	37.5%	18.75	31.25%	12.5%	-
10	43.75%	50%	6.25%	-	-
12	62.5%	31.25%	6.25%	-	-
14	62.5%	6.25%	12.5%	12.5	6.25%

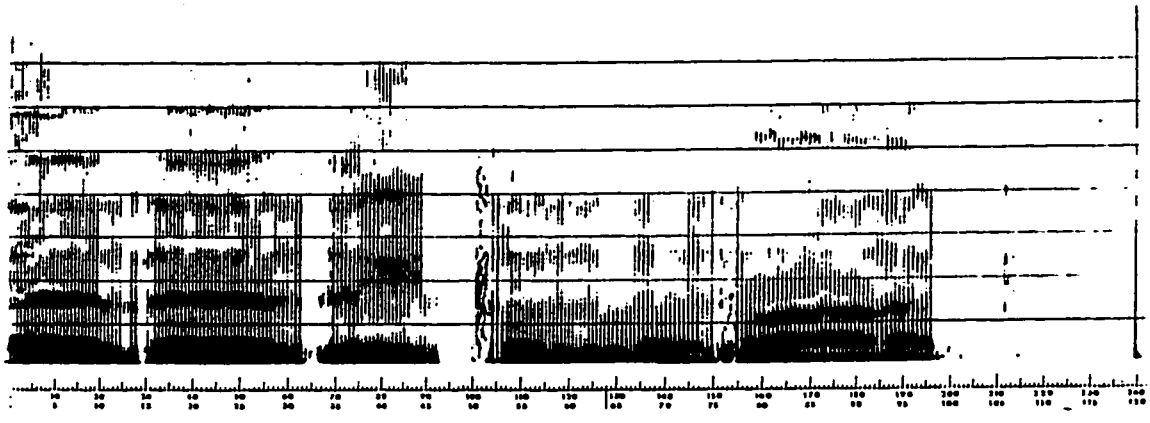
group3

poles number	order				
	1	2	3	4	5
6	75%	12.5%	12.5%	-	-
8	43.75%	43.75	-	12.5%	-
10	25%	18.75%	50%	6.25%	-
12	12.5%	25%	12.5%	43.75	6.25%
14	12.5%	37.5%	18.75%	31.25	-

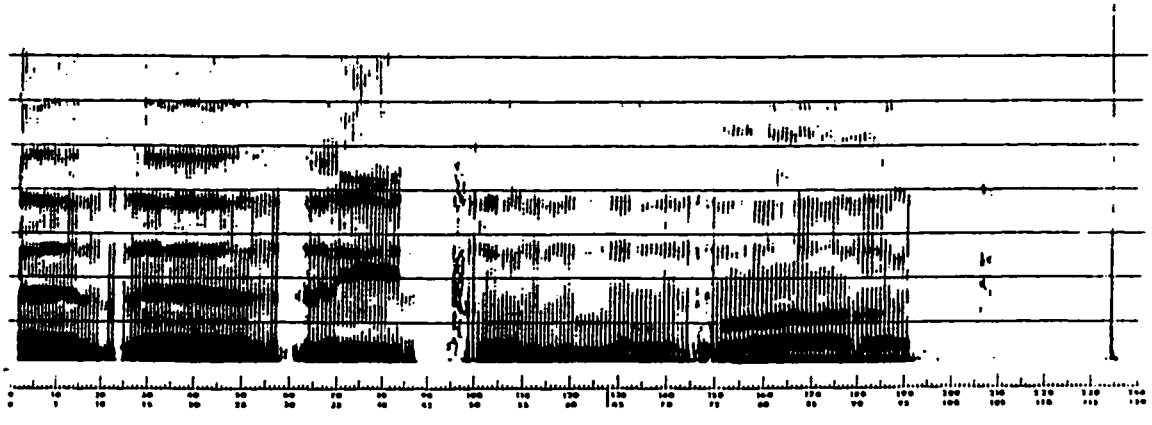
الشارع نظيف



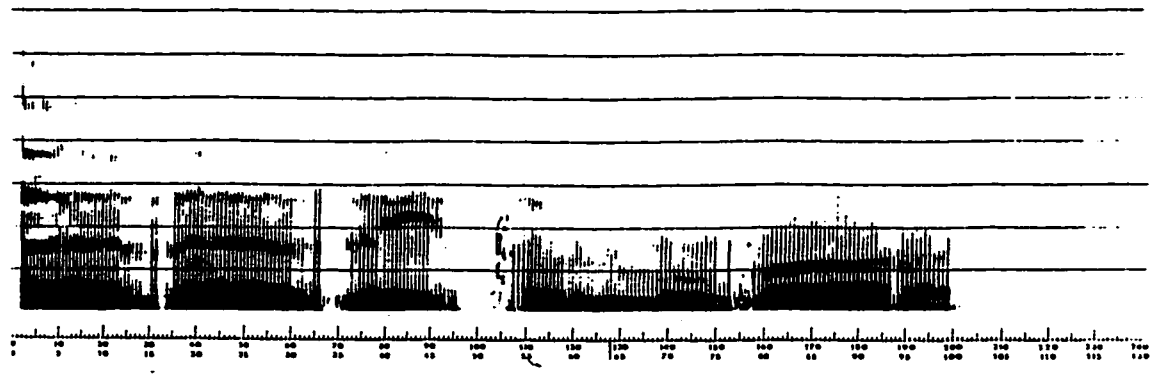
original



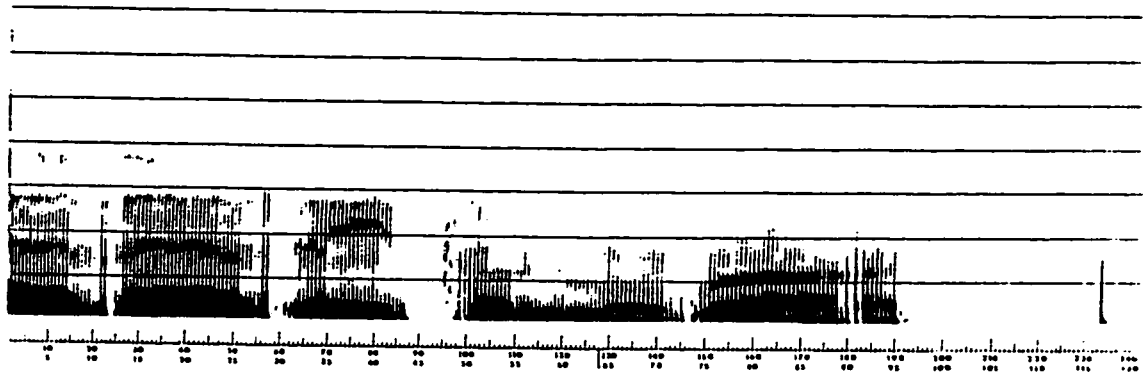
synthesized speech using NP = 6



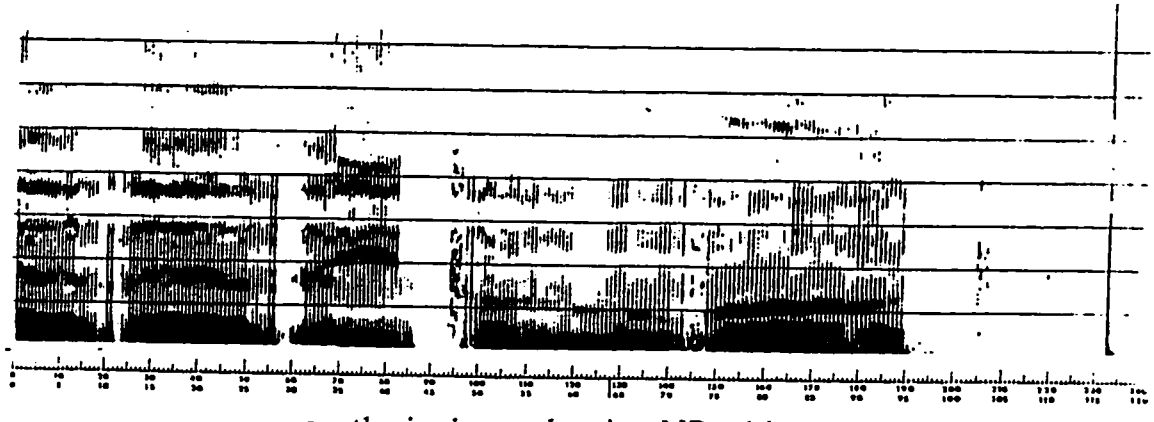
synthesized speech using NP = 8



synthesized speech using NP = 10



synthesized speech using NP = 12



synthesized speech using NP = 14

Figure 4.6: Spectrographs of one sample set of the second test.

4.4.2 Synthesis Parameters

4.4.2.1 Driving Function

Driving function is the most important parameter in the synthesis mode [4]. The noisy driving function for unvoiced frames was kept the same as in Markel and Gray system [14]. However, three different modified driving function for the voiced frames were suggested and they are as follows.

1. $\frac{1}{IPC+1}$

2. $e^{-(IPC+1)/2}$

3. $(-1)^{IPC+1}e^{-(IPC+1)/2}$

Where IPC is a varying parameter which represents pitch period counter. The logic beyond these suggested types of driving functions is as follows. In the LPC model, the prediction error signal is defined as

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

Where e(n) is the error function and s(n) is the speech signal. To the extent that the actual speech signal is generated by the linear prediction system of order p, e(n) is equally a good approximation of the driving (excitation) source. This function is approximated by an impulse train shown in Figure 4.7 for voiced sounds and random noise for unvoiced sounds. This approximation for the error functions deletes some features of it which may affect the smoothness of the synthesized speech. In this thesis modified driving functions are used to resemble the actual error function more closely. In the first two of the modified driving functions shown in Figures 4.8 and 4.9, the objective was to smooth the impulse input by extending its decaying duration to some value to simulate the realistic situation.

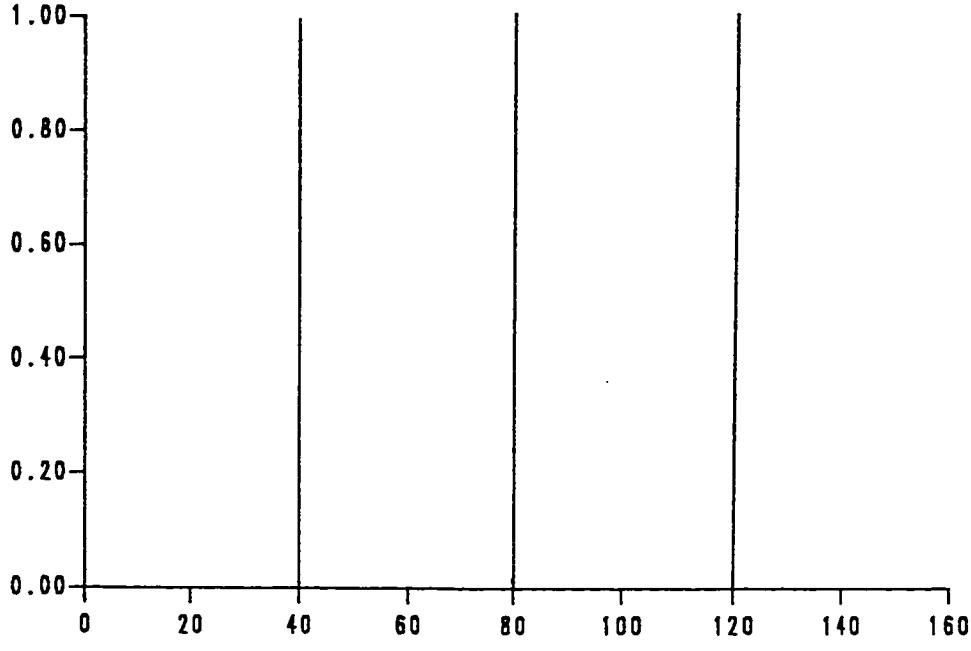


Figure 4.7: Train of impulses driving function

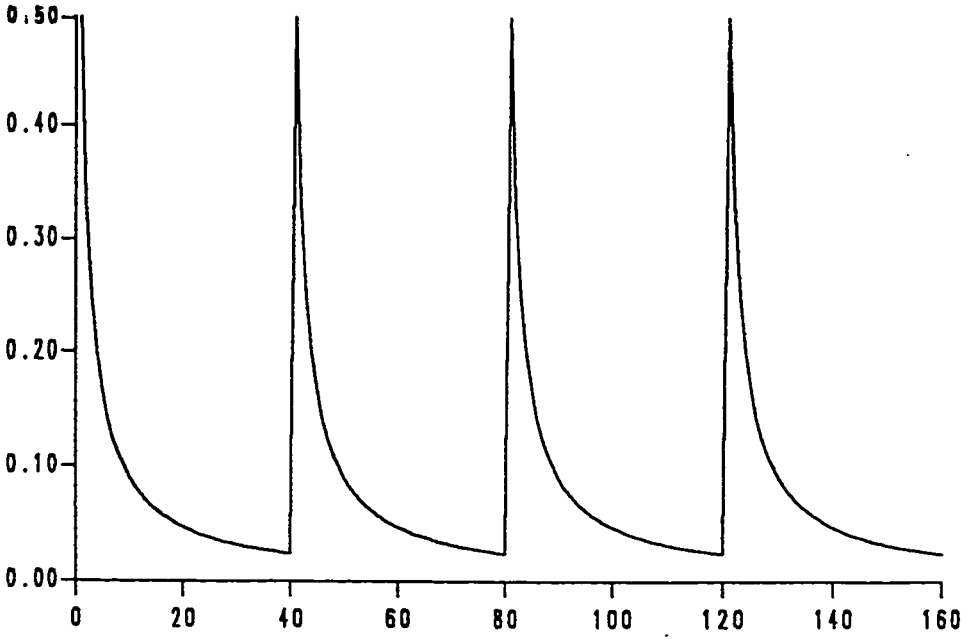


Figure 4.8: First suggested driving function

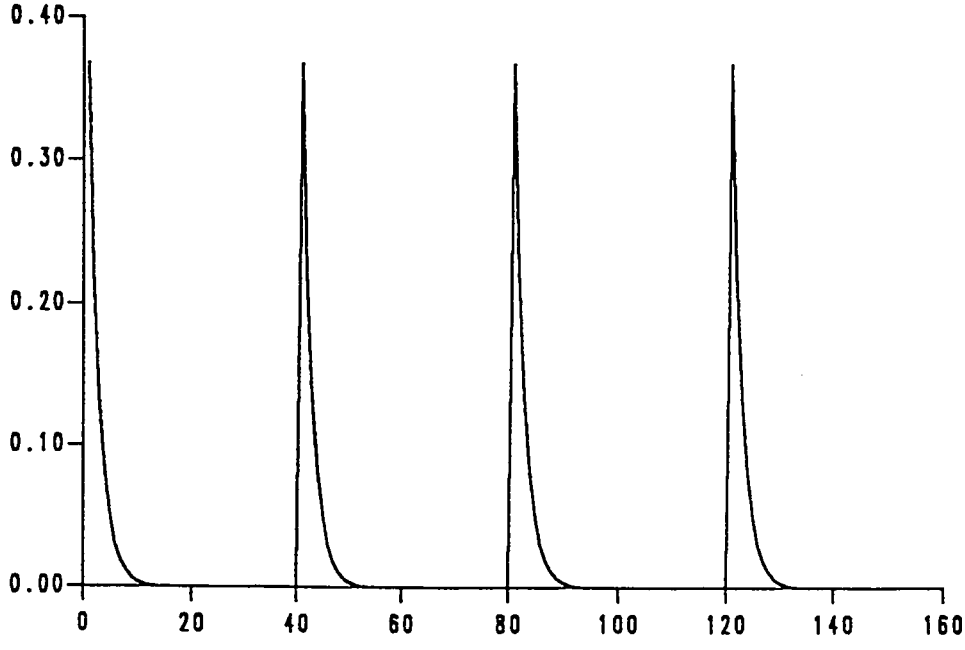


Figure 4.9: Second suggested driving function

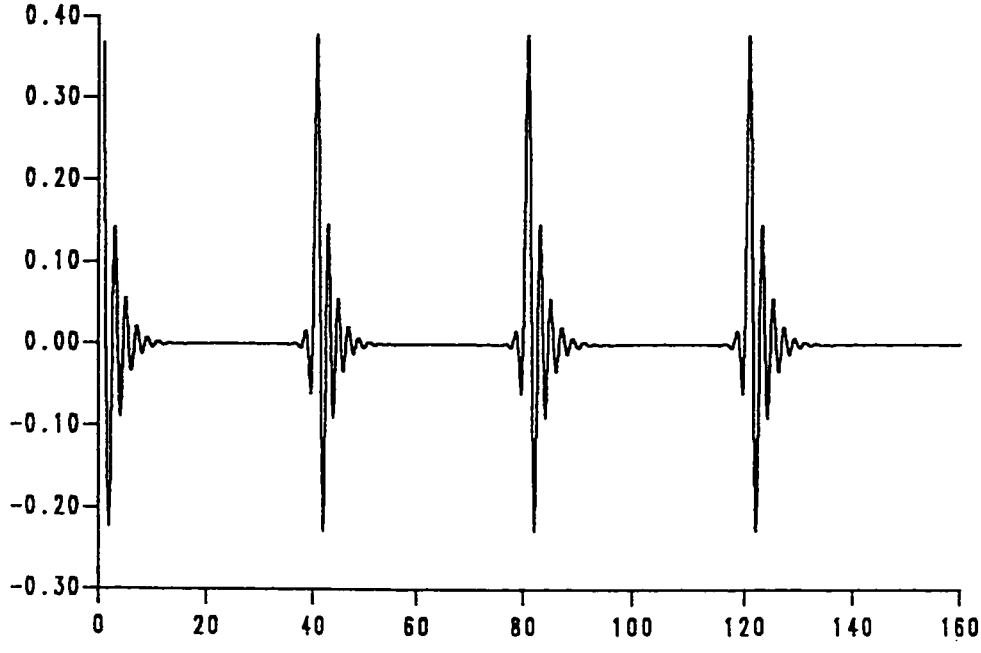


Figure 4.10: Third suggested driving function

In the third modified driving function shown in Figure 4.10, the objective was to switch between positive and negative values of the extented tail duration which makes it closer to the real shape of error function.

In the same manner done for the other tested parameters, three statements form three different speakers were tested. The results of the questionnaire is given in Table 4.3 for this test. Third modified driving function gave the best result which is also confirmed by the spectrographic test. A sample test of the spectrographic test is shown in Figure 4.11.

4.5 The Selected System Parameters

Table 4.4 shows the average percentage of each tested values for each tested parameter. The suitable values of each parameter was selected according to the highest average percentage . The achieved set of suitable parameters for our LPC Arabic systems from the above tests are

1. sampling frequency of 6.5 khz.
2. frame length of 150 samples.
3. frame rate of 43 hz.
4. number of poles equals to 8.
5. preemphasis factor of 0.9 .
6. driving function for voiced speech is a train of $(-1)^{IPC+1}e^{-(IPC+1)/2}$ function and for unvoiced speech a pseudo-random numbers uniformly distributed.

The final result obtained from the spectrograph test are similar to these obtained by the questionnaire test. The spectrographs of an example of synthesized speech and original are shown in Figures 4.12, 4.13 and 4.14 for comparison.

Table 4.3: Driving function testing results

group1

driving function	order				
	1	2	3	4	5
original	100%	-	-	-	-
$\sum_k \delta(t\text{-}IPC_k)$	-	18.75%	37.5%	37.5%	6.25%
$\frac{1}{IPC}$	-	12.5%	43.75%	43.75%	-
$e^{-(IPC+1)/2}$	-	6.25%	12.5%	31.25%	50%
$(-1)^{IPC+1} e^{-(IPC+1)/2}$	-	93.75%	-	6.25%	-

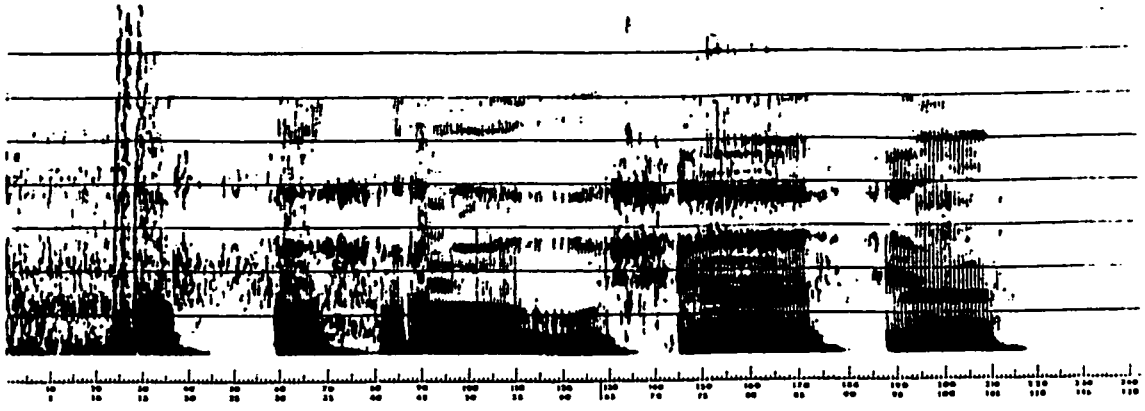
group2

driving function	order				
	1	2	3	4	5
original	100%	-	-	-	-
$\sum_k \delta(t\text{-}IPC_k)$	-	6.25%	31.25%	18.75%	43.75%
$\frac{1}{IPC}$	-	37.5%	18.75%	43.75%	-
$e^{-(IPC+1)/2}$	-	31.25%	43.75%	18.75%	6.25%
$(-1)^{IPC+1} e^{-(IPC+1)/2}$	-	81.25%	12.5%	6.25%	-

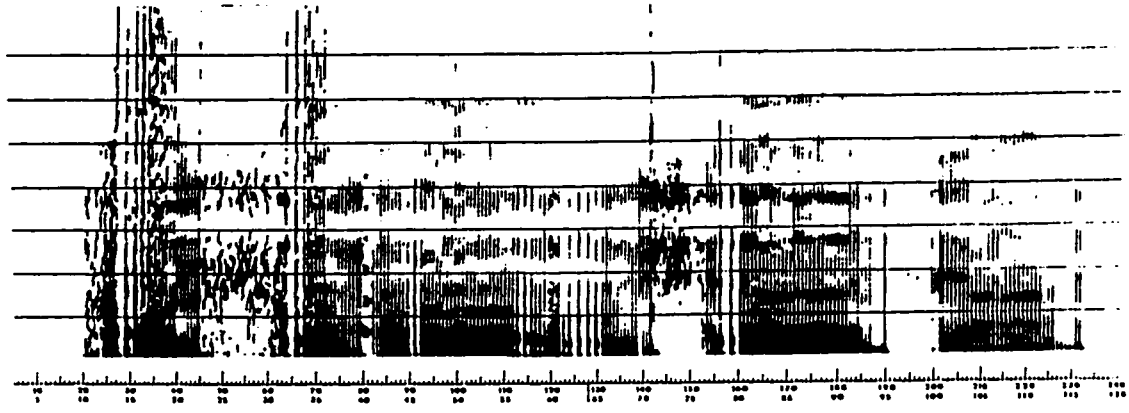
group3

driving function	order				
	1	2	3	4	5
original	87.5%	-	6.25%	6.25%	-
$\sum_k \delta(t\text{-}IPC_k)$	12.5%	43.75%	18.75%	25%	-
$\frac{1}{IPC}$	12.5%	31.25%	31.25%	12.5%	12.5%
$e^{-(IPC+1)/2}$	12.5%	50%	25%	-	12.5%
$(-1)^{IPC+1} e^{-(IPC+1)/2}$	6.25%	37.5%	50%	6.25%	-

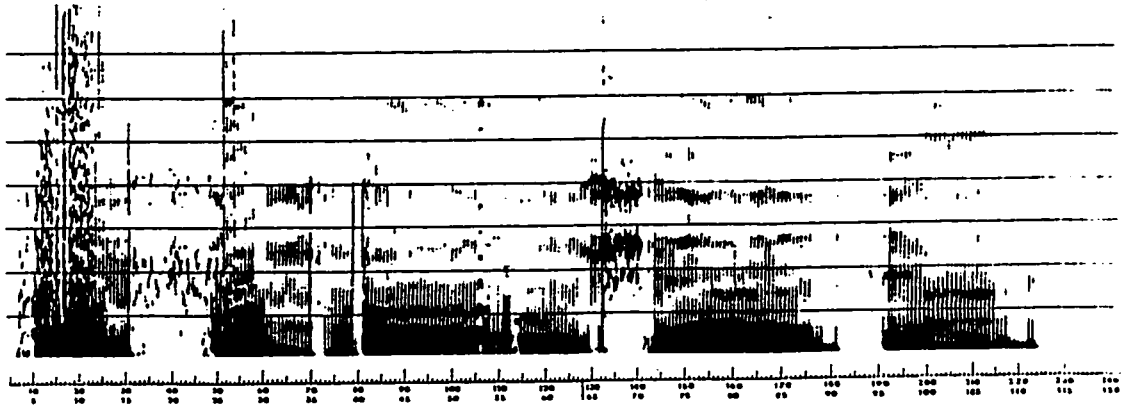
الصحراء شاسعة



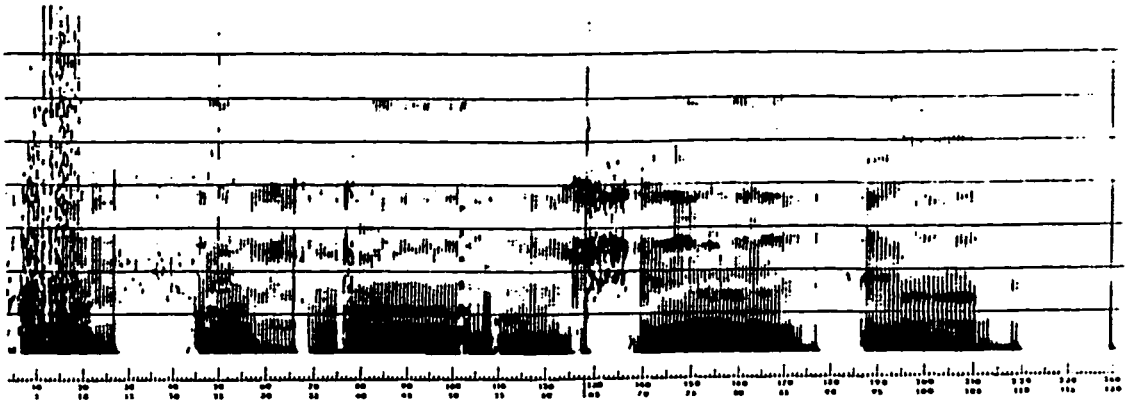
original



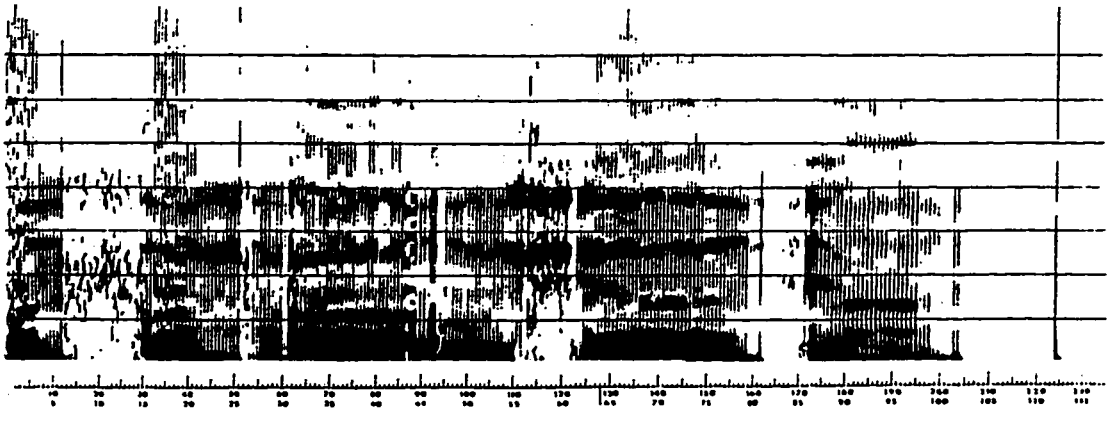
synthesized speech using $\sum_k \delta(t\text{-}IPC_k)$



synthesized speech using $\frac{1}{IPC}$



synthesized speech using $e^{-(IPC+1)}$



synthesized speech using $(-1)^{IPC+1}e^{-(IPC+1)}$

Figure 4.11: Spectrographs of one sample set of the third test.

Table 4.4: The average results of the tested parameters.

Test no. 1

frame length	order				
	1	2	3	4	5
original	92%	6.25%	2.1%	-	-
100	4.17%	-	10.33%	35.42%	45.83%
128	-	16.67%	29.17%	45.83%	8.33%
150	4.17%	79.17%	12.5%	4.17%	-
175	4.17%	16.67%	62.5%	16.67%	-

Test no. 2

poles number	order				
	1	2	3	4	5
6	39.58%	25%	20.83%	8.33%	4.17%
8	50%	31.25	10.42	8.33	-
10	45.83%	27.08%	22.92%	4.17%	-
12	39.58%	25%	16.67%	16.67%	2.08%-
14	39.58%	14.58%	25%	18.75%	2.08%

Test no. 3

driving function	order				
	1	2	3	4	5
original	95.83%	-	2.08%	2.08%	-
$\sum_k \delta(t\text{-}IPC_k)$	4.17%	22.92%	29.17%	27.1%	16.7%
$\frac{1}{IPC}$	4.17%	27.08%	31.25%	33.33%	4.17%
$e^{-(IPC+1)/2}$	4.17%	29.17%	27.08%	16.67%	22.92%
$(-1)^{IPC+1} e^{-(IPC+1)/2}$	2.08%	70.83%	20.83%	6.25%	-

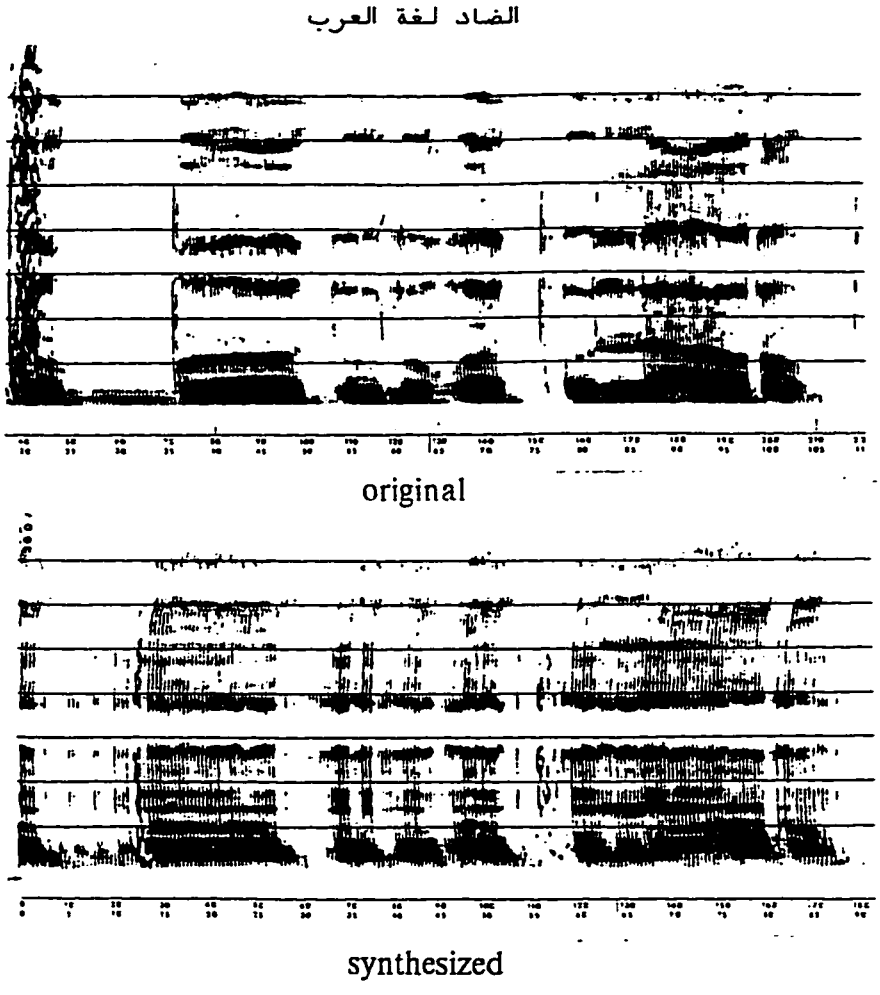


Figure 4.12: Original speech and synthesized speech using the selected parameters.

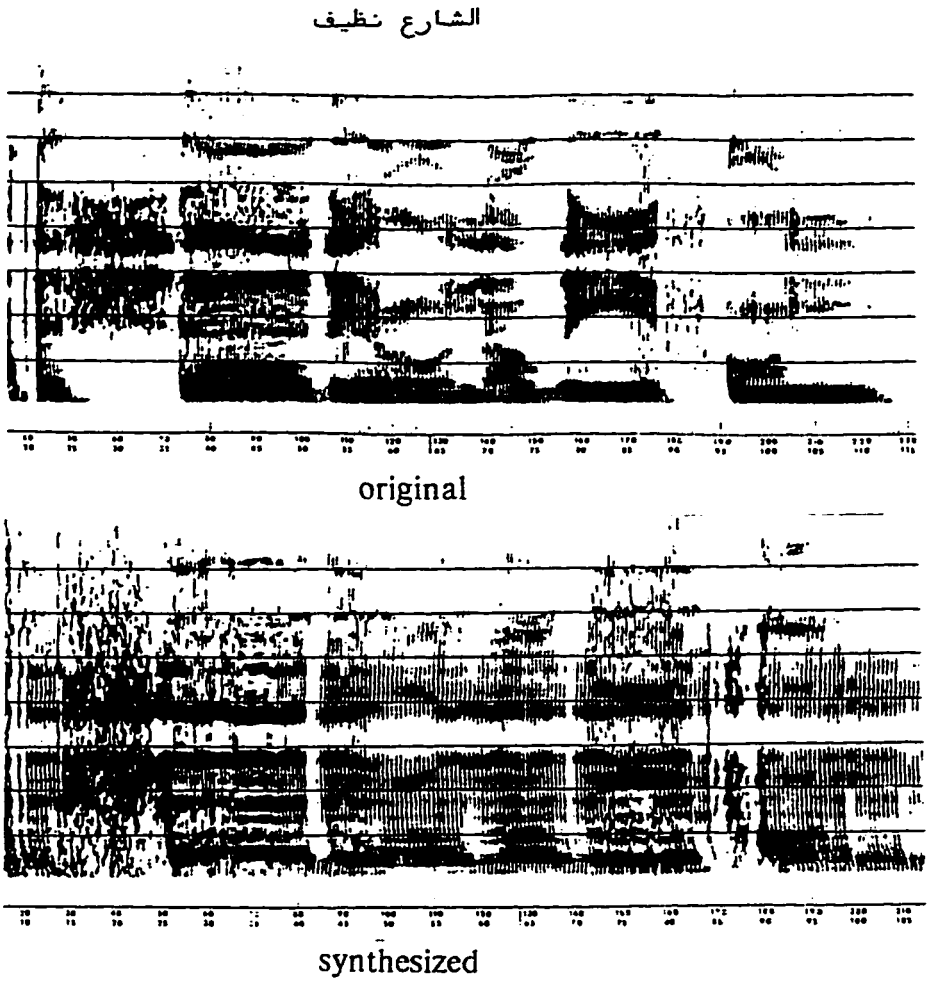


Figure 4.13: Original speech and synthesized speech using the selected parameters.

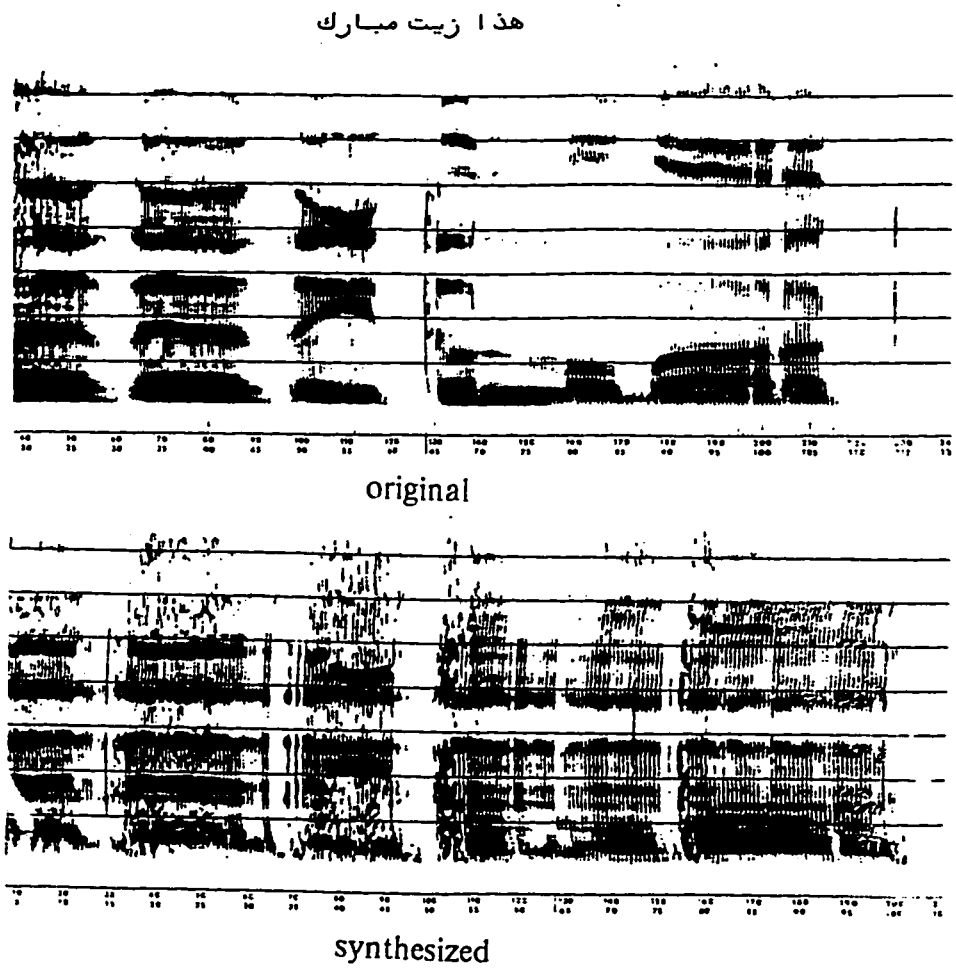


Figure 4.14: Original speech and synthesized speech using the selected parameters.

4.6 Discussion

Eventhough the LPC technique can be used generally for synthesizing any segment of speech of any language, the performance of a particular system using this technique depends on the values of the system parameters which depends on the certain language used. This is clear from the results of this study, where the values of some parameters of the analysis and synthesis system for Arabic were different from these used for English. One of the reasons for these differences is the way in which each language speakers produce the different sounds. For example, English speakers in general speak fast and use the front end of their vocal tract and lips end for speaking. On the other hand, Arabic speakers usually speak slowly and use both the front and back of their vocal tract in producing different sounds. Accordingly, some of the speech signal characteristic for a similar sound would be different. One of these important differences is that the average range of the fundamental frequency F_0 for Arabic vowels is 156-603 hz [12]. The corresponding range for English vowels is 270-730 hz [1]. As a result the analysis frame length used in the system is affected. The largest pitch period for Arabic vowels, in terms of samples at 6.5 khz sampling frequency, is equal to 42 samples. Whereas, for English vowels the largest pitch period is 24 samples. As a result, the analysis frame length for Arabic voiced sounds should be longer, by some amount, than the one used for English voiced sounds to cover enough pitch periods to apply the autocorrelation analysis method for good estimation of the LPC coefficients.

Another reason, for the differences between the LPC system parameters for Arabic and English, is that the existance of some sounds in Arabic language (about 8 sounds) which do not exist in English.

For the over all evaluation of the LPC Arabic analysis and synthesis system, we can say that reasonable results were obtained in spite of the following

1. The trade-off between reducing the transmission and storage requirements when using this technique and the good quality and naturalness of synthesized speech.
2. The effect of the background noise in recording the original speech on degrading the system performance since we did our experiments in a normal laboratory environment which contains some kind of noise. This noise which exists in all practical situations affects the accurate estimation of the pitch period and the evaluation of the LPC coefficients.
3. The classification for each frame as all voiced or all unvoiced in the pitch estimation process. There are obviously speech sounds that would be more accurately classified as something in between. The binary decision is usually made because of practical considerations.
4. The correct and clear pronouncation of the original input sounds to the system has its own effect on the performance of the system in synthesizing clear pronouncation sounds.
5. The class of plosive sounds are very difficult to be accurately analyzed because of their small durations which are often less than or close to the analysis frame.
6. Finally the assumption of the driving (excitation) function for the synthesizer which may not be good approximation of the error function for all types of utterances and all speakers.

Chapter V
CONCLUSION

5.1 Summary

The discussion so far demonstrated that eventhough the LPC technique can be used for implementing a speech analysis and synthesis system for any language, the conditions of the system parameters for one language is different from those of another language.

In this thesis work a reported speech analysis and synthesis system in the literature [14] was implemented and tested for Arabic language. The obtained synthesized Arabic speech was not satisfactory. Accordingly the system is remodeled to produce good quality Arabic speech. The reported system parameters values are

1. Sampling rate of 6.5 khz.
2. Analysis frame length of 128 samples.
3. Frame rate of 51 hz.
4. Number of poles equals 10.
5. Pre-emphasis factor of 0.9.
6. Driving function for the synthesizer is periodic impulses for voiced frames and pseudo-random numbers for unvoiced frames.

The achieved set of suitable parameters for our LPC Arabic system are

1. Sampling frequency of 6.5 khz.
2. Analysis frame length of 150 samples.
3. Frame rate of 43 hz.

4. Number of poles equals to 8.

5. Preemphasis factor of 0.9.

6. Driving function for voiced speech is a train of $(-1)^{IPC+1}e^{-(IPC+1)/2}$ function and for unvoiced speech a pseudo-random numbers.

In this thesis we also proposed three types of driving (excitation) functions for voiced sounds which gave better results than the proposed one in the literature. The most important system parameter is the frame length as it results in a considerable differances between the original speech and the synthesized speech if the recommended value is not used. The driving function has its effect for producing more natural sounding synthesized speech. By comparing the bit rate of the two systems, based upon a resolution of 12bits/sample and sampling rate of 6.5 khz, the reported system has bit rate of $B=(10*12+1*12+2)*51=6834$ bits/seconds and the implemented system has bit rate of $B=(8*12+1*12+2)*43=4730$ bits/second. The bit rate of a speech signal digitized directly at 6.5 khz is 78000 bits/second. Then the data compression ratio of 16.5 is obtained.

5.2 Recommendation for Further Work

We recommend the set-up of a digital speech processing laboratory with sufficient facilities. In the following, further work in the area of Arabic speech synthesis is presented.

1. Using our developed system in forming a dictionary of smallest elements of Arabic speech sounds, like phonemes, allophones, etc. , by representing these elements by the LPC coefficients and finding a way for concatenating them according to Arabic grammer rules to build a text-to-speech system.

2. Pitch detection forms a major parameter in the analysis and synthesis system. None of the existing pitch detection algorithms can be said to work perfectly. Finding a pitch detection algorithm specially designed for Arabic may improve the quality of the synthesized speech.

3. The binary voiced/unvoiced decision for each analysis frame is not suitable for all frames because some of them can be defined as inbetween. A study of pitch detection algorithm which includes these cases can be done to have a definition of these mixed voiced and unvoiced frames. According to this a new driving function for mixed decided frames should be found.

4. More investigation can be done on the type of the driving function for the synthesizer to find more closer driving function to the prediction error shape. The output of this study is more natural sounding of the synthesized speech.

5. Vector quantaization can be used to code the LPC parameters of our system for more reduction of transmission and storage requirements.

Appendix A
PROGRAMS

A.1 Main program

```
C
      DIMENSION S(0:30000),SN(-1:700),SNW(0:700),A(19),RC(19)
      DIMENSION WN(0:700),SNP(0:700),R(0:21),PBUF(0:300),PITCH(3)
      DIMENSION POTCH(-1:7000),PRE(0:7000),U1(300)
      DIMENSION POTC(-1:7000),AL(7000),PRE1(0:7000),RCB(8000)
      DIMENSION ST(0:700),STN(0:700),STW(0:700)
     INTEGER  DAT(0:30000),X,POTCH
C
        READ(9,*)NS,LF,NP,IW
C
        WRITE(6,*)NS,LF,NP
        NX=0
      PRE(0)=0
      ML=LF/2
C
C     NP NUMBER OF POLES
C     LF LENGTH OF ANALYSIS FRAME
C     NS NUMBER OF SAMPLES
C     IW TO CONTROL ANALYSIS FRAME LENGTH
C     ML FRAME SHIFT
C
      OF=NS-LF
      NF=(OF/ML)+IW-1
      NFE=NF*IW
C
C
```

```
C N11 TOTAL NUMBER OF LPC COEFF.
          N11 = NFE*NP
C
        WRITE(6,*) 'N11',N11,'NF',NF,'LF',LF,NFE
C
C-----------------------------------------------------------------
C GETTING SAMPLED VALUES
C
       READ(7,2)(DAT(I),I = 0,NS-1)
 2   FORMAT(10I7)
       X = 0
       DO 110 J = 0,NS-1
       S(J) = DAT(J)/204.7
  110  CONTINUE
C-----------------------------------------------------------------
C
C  FIND WINDOW VALUES
C
       LFE = LF/IW
       C = 2*3.141592/(LFE-1)
       DO 10 I = 0,LFE-1
       WN(I) = 0.54 -0.46*COS(C*I)
 10   CONTINUE
C-----------------------------------------------------------------
C
C----------------------ANALYSIS MODE-------------------------------
C
```

```
C  TAKE ONE FRAME
        IY=1
        IQ=1
C
        DO 20 K=1,NF
        MX=ML*(K-1)
        DO 30 N=0,LF-1
        SN(N)=S(MX+N)
  30  CONTINUE
C-----------------------------------------------------------------
C  TO FIND THE VALUE OF A1 IN THE PRE-EMPASES PART OF THE PROCESS
C
        SUM1=0.
        DO 40 KK=0,LF-2
        SUM1=SUM1+SN(KK)*SN(KK+1)
  40  CONTINUE
        SUM2=0
        DO 50 KI=0,LF-1
        SUM2=SUM2+SN(KI)**2
  50  CONTINUE
        U=   SUM1/SUM2
        IF(U.LT.0.6)A1=0
        IF(U.GE.0.6)A1=0.9
C-----------------------------------------------------------------
C   PRE-EMPASIZING THE DATA
C
         PRE(K)=A1
```

```
      SN(-1)=0.
      DO 60 J=0,LF-1
      SNP(J)=SN(J)-A1*SN(J-1)
   60 CONTINUE
C ---------------------------------------------------------------
C CALCULATING LPC FOR THE CURRENT FRAME K BY TAKING IW SUB-FR
C
      MX1=0
      DO 140 I=1,IW
      DO 150 IU=0,LFE-1
      STN(IU)=SNP(IU+MX1)
  150 CONTINUE
      MX1=MX1+LFE
C---------------------------------------------------------------
C  WINDOWING THE FRAME
C
      DO 170 N=0,LFE-1
      STW(N)=WN(N)*STN(N)
  170 CONTINUE
C---------------------------------------------------------------
C EVALUATING THE LPC COEFFECIENTS
C
        CALL AUTO(LFE,STW,NP,A,ALPHA,RC)
C
        AL(IY)=SQRT(ALPHA)
C
        NX1=NP*NX
```

```
            DO 180 NG=1,NP
            RCB(NG+NX1)=RC(NG)
 180    CONTINUE
C-------------------------------------------------------------
            NX=NX+1
C
C -------------------------------------------------------------
            IY=IY+1
C
 140    CONTINUE
 20     CONTINUE
C PRE-EMPHASIS VECTOR
            DO 66 IK=1,NF
            DO 66 IH=0,IW-1
            PRE1(IQ)=PRE(IK)
            IQ=IQ+1
 66     CONTINUE
C
C
            READ(11,114)(POTCH(I),I=1,NFE)
 114    FORMAT(1514)
C
C-------------------------------------------------------------
C
C-------------------------SYNTHESIS MODE----------------------
C
            CALL SYNTW(RCB,POTCH,AL,PRE1,N11,NFE,LFE,NP,FS)
```

```
C
      STOP
      END
```

A.2 Autocorrelation Subroutine for Calculating The LPC's Coefficients

```
C ---------------------------------------------------------------
C SUBROUTINE AUTO
C A SUBROUTINE FOR IMPLEMENTATION THE AUTOCORRELATION MTHO
C PREDICTION ANALYSIS.
C ---------------------------------------------------------------
C
      SUBROUTINE AUTO(N,X,M,A,ALPHA,RC)
C
C  INPUTS:  N - NO. OF DATA POINTS
C           X(N)- INPUT DATA SEQUENCE
C           M- ORDER OF FILTER (M <21,)
C  OUTPUTS: A- FILTER COEFFICIENTS
C           ALP- RESIDUAL "ENERGY"
C           RC - "GENERATED REFLECTION COEFFICIENTS"
C
C  PROGRAM LIMITED TO M <21 , BECAUSE OF DIMENTION OF R(.)
C
C   SUBROUTINE AUTO(N,X,M,A,ALPHA,RC)
C
      DIMENSION X(0:N),A(M + 1),RC(M + 1)
      DIMENSION R(0:21)
       MP = M
      DO 10 K = 0,MP
      R(K) = 0.
      K1 = K + 1
      NK = N-K1
```

```
      DO 10 NM=0,NK
 10   R(K)=R(K) + X(NM)*X(NM+K)
C     Y=R(0)
C     DO 15 I=0,MP
C     R(I)=R(I)/Y
C15   CONTINUE
      RC(1)=-R(1)/R(0)
      A(1)=1.
      A(2)=RC(1)
      ALPHA= R(0)+R(1)*RC(1)
      DO 40 MINC=2,M
      S=0.
      DO 20 IP=1,MINC
 20   S=S+R(MINC-IP+1)*A(IP)
      RC(MINC)=-S/ALPHA
      MH=MINC/2 +1
      DO 30 IP=2,MH
      IB=MINC-IP+2
      AT=A(IP)+RC(MINC)*A(IB)
      A(IB)=A(IB)+RC(MINC)*A(IP)
 30   A(IP)=AT
      A(MINC+1)=RC(MINC)
      ALPHA=ALPHA +RC(MINC)*S
      IF(ALPHA) 50, 50, 40
 40   CONTINUE
 50   RETURN
      END
```

A.3 AMDF Pitch Detection Program

```
C
      DIMENSION S(0:30000),DAT(0:30000),SN(-1:700),SNP(0:700)
      DIMENSION D(0:700),SNP1(-300:300),PRE(1000)
     INTEGER  DAT,T(0:700),TI(0:700),PITCH(0:1000)
C
      READ(9,*)NS,K0
      WRITE(6,*)NS,K0
      ML=K0/2
      IW=2
C
      OF=NS-K0
C     NF=(OF/ML)+10
      NF=(OF/ML)+IW-1
C
C-----------------------------------------------------------------
C GETTING SAMPLED VALUES
C
      READ(7,2)(DAT(I),I=0,NS-1)
 2   FORMAT(10I7)
      X= 0
C     FILTER WITH CUTOFF 900HZ,6500
      DATA P0,P1,P2,P3/.0704583,.01374714,.01374713,.07045829/
      DATA A0,A1,A2,A3/1.,-1.273449,0.9377728,-.2484641/
C
C  LOW PASS FILTER SPEECH SAMPLES
C
```

```
      SOUT=0
      Y1=0
      Y2=0
      Y3=0
C
      X1=0
      X2=0
      X3=0
C
      DO 110 J=0,NS-1
      X0=DAT(J)
      SOUT=P0*X0+P1*X1+P2*X2+P3*X3-A1*Y1-A2*Y2-A3*Y3
C
C
      Y3=Y2
      Y2=Y1
      Y1=SOUT
C
      X3=X2
      X2=X1
      X1=X0
C     IF (MOD(J,II).NE.0) GO TO 20
C     K=J/II
C     PBUF(K)=SOUT
      S(J)=SOUT/204.7
  110 CONTINUE
C-----------------------------------------------------------------
```

```
C
C
C  TAKE ONE FRAME
C
      DO 100 K=1,NF
      MX=ML*(K-1)
      DO 10 N=0,K0-1
      SN(N)=S(MX+N)
   10 CONTINUE
C-----------------------------------------------------------------
C                  FIND THE DIFFERENCE FUNCTION
       KY=K0/2
C
      DO 62 KI=0,K0-1
       SNP1(-KY+KI)=SN(KI)
C     SNP1(-KY+KI)=SNP(KI)
   62 CONTINUE
C     WRITE(6,*)'KI',KI
C-----------------------------------------------------------------
      DO 90 I=0,K0/2
        SUM=0
       DO 80 J=1,K0/2-1
        SUM=SUM+ABS(SNP1(J)-SNP1(J-I))
   80 CONTINUE
        D(I)=SUM/KY
C
   90  CONTINUE
```

```
C
C                    DECISION
C
        AMIN=1000
       DO 70 J=10,K0/2-1
        IF(D(J).GT.AMIN)GO TO 70
          AMIN=D(J)
          T(K)=J
   70   CONTINUE
C
        AMAX=-1000
        DO 71 JJ=1,K0/2-1
        IF(D(JJ).LT.AMAX)GO TO 71
         AMAX=D(JJ)
   71   CONTINUE
C
          IJK=0
          DES=AMAX-AMIN
C      IF(DES.GT.0.2)IJK=1
          IF(DES.GT.0.1)IJK=1
          IF(IJK.EQ.1.)GO TO 72
          T(K)=0
   72   CONTINUE
C
  100  CONTINUE
       WRITE(6,*)MX+N
        DO 61 L=1,NF
```

```
C
      TI(L)=T(L)
 61   CONTINUE
      IQ=1
      DO 64 I=1,NF
      DO 64 KW=0,IW-1
      PITCH(IQ)=TI(I)
      IQ=IQ+1
 64   CONTINUE
      WRITE(11,34)(TI(K),K=1,NF)
      WRITE(11,*)'
C     WRITE(11,32)(PRE(K),K=1,NF)
C     WRITE(8,33)(PITCH(K),K=1,NF)
      WRITE(8,33)(PITCH(K),K=1,IW*NF)
 33   FORMAT(15I4)
 34   FORMAT(10I4)
 32   FORMAT(10F3.2)
C
      STOP
      END
C
```

A.4 Synthesis Subroutine

```
C-----------------------------------------------------------------
C
C  SYNTHESIS SUBROUTINE TO FIND SYNTHESIZED SPEECH
C
C-----------------------------------------------------------------
C
      SUBROUTINE SYNTW(RCB,POTCH,AL,PRE,N11,N22,LF,NP,FS)
C
      DIMENSION RCB(8000),AL(7000),POTCH(-1:7000),PRE(0:7000)
      DIMENSION RCL(20),RCR(20),RC(20),RCBUF(20),Y(400),TAP(20)
      DIMENSION P(20)
      INTEGER POTCH
C
      REAL NOISE(30000)
      INTEGER NR,IY(400)
C
      FR=.2
      FL=.8
      IT=0
      PRE(0)=0
      WRITE(6,*)'LF',LF
C.....NUMBER OF OUTPUT SAMPLES.....
      ND11=N22*LF/2
      IF(LF.EQ.150)ND11=N22*LF/2 -150/2
      WRITE(8,*)ND11
C.....RANDOM NUMBER GENERATOR.....
```

```
C
      NR=30000
C
      READ(4,33)(NOISE(I),I=1,30000)
   33 FORMAT(10E12.5)
C
      DO 13 I=1,30000
   13 NOISE(I)=(NOISE(I)-.5)*.2
C .....
C
C.....
C
      N=LF
      N0=N/2
      IFLGTH=N/2
C
      DATA RC,RCR,RCBUF,TAP/80*0/
      DATA P/1,19*0/
      DATA PR,DRV/0.95,0/
      DATA IFC,IVR,NN,GAIN/1,0,1,0/
      DATA YPREV,IRN,NB/0.,8,1/
      DATA IPTCHL,IVL,DRVN,GAINR/0,1,2*0./
C
C
      IPTCHR=N/2
      IPITCH=N/2
      IPC=1
```

```
C
C
      LSTBLK=N22
C.....
C
  10  IF(IPC.LE.IPITCH)GO TO 120
C
C
C---------- STARTING A NEW PITCH PERIOD ----------
C-------------------------------------------------------------
C
  20  IPC=1
  30  CONTINUE
      IF(IFC.LE.IFLGTH)GO TO 80
C
C---------- CROSSING ANALISIS FRAME BOUNDRY----------
C-------------------------------------------------------------
C
C
      IFC=IFC-IFLGTH
C
      IF(NB.GT.LSTBLK)GO TO 140
      DO 40 J=1,NP
      RCL(J)=RCR(J)
  40  CONTINUE
      IPTCHL=IPTCHR
      GAINL=GAINR
```

```
C
      IVLST=IVL
      IVL=IVR
      IF((IVLST.EQ.0).OR.(IVL.EQ.1)) GO TO 60
C
C ZERO BUFFER DURING TRANSITION FROM V. TO UV.
C
      DO 50 J=1,NP
   50 RCBUF(J)=0.
   60 CONTINUE
C
C                  ----------
C---------------------  READ NEW PARAMETERS -------------------------
C              RCR(1),.....,RCR(M),SIGMA,IPTCHR
C             ---------------------------------=
C
      MK=NP*(NB-1)
      DO 2 JN=1,NP
      RCR(JN)=RCB(MK+JN)
 2    CONTINUE
C
      SIGMA=AL(NB)
      PR=PRE(NB)
C
C
      IPTCHR=POTCH(NB)
C
```

```
C
      NB=NB+1
      IF(IPTCHR.GT.0) GO TO 70
C
C    THE GAIN FOR NOISE OR UNVOICED FRAME
C
      GAINR=SIGMA*SQRT(3./N)*10
C
C
      IVR=0
      GO TO 30
C    THE GAIN FOR VOICED FRAME
C
   70 GAINR= SIGMA/SQRT(FLOAT(N))
      IVR=1
C
C---------- END OF CROSSING FRAME BOUNDARY WORK.----------
C
C---------- PERFORM INTERPOLATION ----------
C-------------------------------------------------------------
C
   80 IF((IVR.EQ.1).AND.(IVL.EQ.1)) GO TO 90
      IPITCH=IPTCHL
      IF (IVL.EQ.0) IPITCH=IFLGTH-IFC+1
      GO TO 100
   90 CONTINUE
C
```

```
 38   IPITCH = FR*IPTCHR + FL*IPTCHL
C
 100  DRV = 1
C
      DRVN = -1/(IPITCH-1)
C
      DO 110 J = 1,NP
C
 110  RC(J) = FR*RCR(J) + FL*RCL(J)
      GAIN  = FR*GAINR + FL*GAINL
C
      IF(IVL.EQ.1) GAIN = GAIN*SQRT(FLOAT(IPITCH))
C
C
C------------------ END OF SETUP FOR A NEW PITCH PERIOD ----------
C ---------------------------------------------------------------
C
 120  IF(IVL.EQ.1) GO TO 130
C
C
      DRV = NOISE(IRN)
C
C
 130  TAP(1) = GAIN
C-----------------------SYNTHESIS ---------------------------------
C-----------------------------------------------------------------
C
```

```
      CALL TWOMUL(RC,TAP,NP,RCBUF,DRV,YOUT)
C
      Y(NN)=YOUT+PR*YPREV
      DRV=DRVN
C.....
      YPREV=Y(NN)
      IFC=IFC+1
      NN=NN+1
      IPC=IPC+1
      IRN=IRN+1
      IF(IRN.GE.NR)IRN=2
      IF(NN.LE.N0) GO TO 10
      NN=1
C
C---------- WRITE OUT N0 SYNTHESIZED SAMPLES ----------
C ----------
C
      DO 12 II=1,N0
      IY(II)=Y(II)*204.7
   12 CONTINUE
C
      IT=IT+1
      IF (IT.GT.2)WRITE(8,3)(IY(IW),IW=1,N0)
    3 FORMAT(1515)
      GO TO 10
C
  140 WRITE(9,*)IT
```

```
      RETURN
      END
C*************************************************************
C  TWO MULTIPLIER LATTICE SYNTHESIS MODEL
C
C*************************************************************
C
      SUBROUTINE TWOMUL(RC,TAP,M,EM,EP,XOUT)
C
      DIMENSION RC(20),TAP(20),EM(20)
      XOUT=0
      DO 10 I=1,M
      II=M+1-I
      JJ=II+1
      EP=EP-RC(II)*EM(II)
      EM(II+1)=EM(II)+RC(II)*EP
   10 XOUT=XOUT+EM(II+1)*TAP(II+1)
      EM(1)=EP
      XOUT=XOUT+EM(1)*TAP(1)
      RETURN
      END
C
```

Appendix B
SPECIFICATION OF THE EQUIPMENTS

Amplifier

Teac A-3440 Four Channel Tape Deck

Microphone

AKAI Dynamic Microphone ADM-80

Filters

-General Radio Copmany 1952 Univirsal Filter.

-General Radio Company 1925 Multifilter One Third Octave 25hz-20khz.

A/D Converter

Data Translation DT2801-A Board

(installed on COMPAC DESKPRO 386 PC)

Resolution of 12 bits/sample

Input Range of -10 to +10 volts.

Sampling frequency is Program contrallable maximum of 33khz

REFERENCES

1. L. R. Rabiner & R. W. Schafer, ***Digital Processing of Speech Signal***,Prentice Hall 1978.
2. R. Linggard, ***Electronic Synthesis of Speech***,Cambridge University Press, 1985.
3. R. Descout, "Speech Synthesis", ***Arabic School of Science and Technology: Applied Arabic Linguistics and Signal and Information Processing, (proceedings)***, Morocco 1983.
4. J. D. Markel & Jr. A. H. Gray, ***Linear Prediction of Speech***, Springer-Verlag, New York, 1976.
5. F. Itakura & S. Satio, "Analysis Synthesis Telephony Based on The Maximum Likelihood Method", ***The 6th International Congress on Acoustics, Tokyo, Japan***, August 21-28, 1968.
6. B. S. Atal & S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction", ***J. Acoust. Soc. Amer.***, Vol.50, pp. 637-655, August 1971.
7. J. Makhoul, "Linear Prediction: A Tutorial Review", ***IEEE Proc.***, pp. 561-580, April 1975.
8. J. Makhoul, S. Rocos & H. Gish, "Vector Quantization in Speech Coding", ***IEEE Proc.***, Vol. 73, No. 11, pp. 1551-1588, Nov., 1985.
9. Y. El-imam, "A Personal Computer-Based Speech Analysis and Synthesis System", ***IEEE Micro***, June 1987, pp.4-21.
10. H. Wakita, "direct Estimation of The vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", ***J. Acoust. Soc. Am.***, vol. 55, pp. 1070-1075, May 1974.

11. P. Ladefoged, R. Harshman, L. Goldstein & L. Rice, "Generating Vocal Tract Shapes From Formant Frequencies", *J. Acoustic Soc. Am.*, Vol. 64, pp. 1027-1035, 1978.

12. S. Penbeci & J. Hejres, "Properties of Arabic Speech Signals for Human-Computer Communication" *The Ninth National Computer Conference & Exhibition Proceedings*, Vol. 2, pp. 12-1-1/12-1-21, Riyadh 1986.

13. Y. A. El-Imam & A. Dannan, The Phonetics of Modern Standard Arabic", *IBM Kuwait Scientific Center*, Ksc 022, Oct. 1986.

14. J. D. Markel & Jr. A. H. Gray, "A Linear Prediction Vocoder Simulation Based upon The Autocorrelation Method", *IEEE Trans.* ASSP, Vol. ASSP-22, pp. 124-134, April 1974.

15. M. Fikri, "A PC-Based Speech Research Facility", *The Ninth National Computer Conference & Exhibition Proceedings*, Vol. 2, pp. 12-4-1/12-4-8, Riyadh 1986.

16. S. Ahasi, "View on The Components of Advanced Arabic Applications", *Computer Processing of Arabic Language - Workshop papers-* Vol.1, Kuwait 1985.

17. A. Mouradi, A. Rajouani & M. Najim, "Text to Speech Conversion in Arabic Language", *Computer Processing of Arabic Language - Workshop Papers - Vol.1, 1985, Kuwait.*

18. J. L. Flanagan, *Speech Analysis, Synthesis and Perception* , Springer-Verlag, New York, 1972.

19. A. M. Noll, "Cepstrum Pitch Determination", *J. Acoust. Soc. Amer.* , Vol. 41, Feb. 1967, pp. 293-309.

20. J. Makhoul, "Spectral Analysis of Speech by Linear Prediction", *IEEE Trans.* Vol. AU-21, pp 140-148, 1973.

21. W.A. Lea, *Trends in Speech Recognition*,Prentice Hall 1980.

22. C.H. Coker, "A Model of Articulatory Dynamics and Control" ***Proceeding of IEEE,*** Vol. 64, NO. 4 ,pp. 452-459, Apr. 1976.

23. L. R. Rabiner & S. E. Levinson, "Isolated and Connected Work Recognition Theory and Selected Applications ***IEEE Trans. Comm.*** , Vol. COM-29, NO. 5 ,pp. 621-659, May. 1981.

24. Walfgang J. Hess, "Algorithms and Devices for Pitch Determination of Speech Signals", ***Proc. of The NATO Advanced Study Institute,*** Bonas France, June 29-July 10, 1981.

25. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg & C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", ***IEEE Trans. ASSP,*** Vol. ASSP-24, No. 5, PP. 399-418, Oct. 1976.

26. C. A. McGonegal, L. R. Rabiner, & A. E. Rosenberg "A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech" ***IEEE Trans. ASSP,*** Vol. ASSP-25, No. 3, PP. 221-229, June 1977.

27. C. A. McGonegal, L. R. Rabiner, & A. E. Rosenberg "A Semiautomatic Pitch Detector (SAPD)", ***IEEE Trans. ASSP,*** Vol. ASSP-23, No. 6, PP. 570-574, Dec. 1975.

28. J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", ***IEEE Trans. Audio Electroacoust.,*** Vol. AU-20, pp. 367-377, Dec. 1972.

29. M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg & H. J. Manley, "Average Magnitude Difference Function Pitch Extractor", ***IEEE Trans. ASSP,*** Vol. ASSP-22, pp. 353-362, Oct. 1974.

30. J. D. Wise, J. R. Caprio & T. W. Parks, "Maximum Likelihood Pitch Estimation", ***IEEE Trans. ASSP,*** Vol. ASSP-24, No. 5, pp. 418-423, Oct. 1976.

31. A. V. Oppenhaeim & R. W. Schafer, *"Digital Signal Proceesing"*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

32. Dr. Hasan A. Qarawi, Privet Discussions.

33. The ILS Software Package, Signal Technology Inc., 5951 Encina Rd., Goleta, CA 93117.